

# HITS

Heidelberg Institute for  
Theoretical Studies

2020

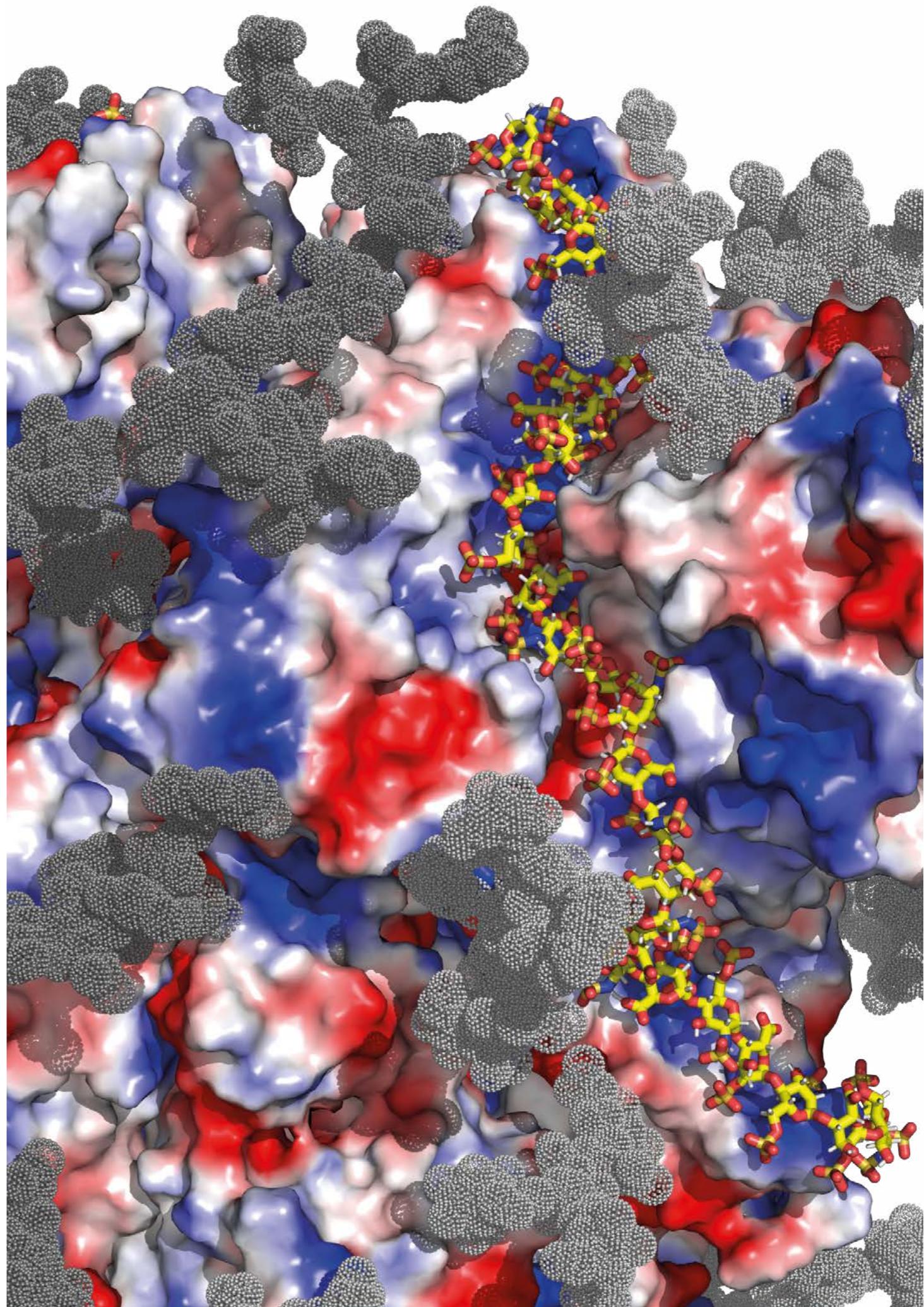
Annual Report  
Jahresbericht

# Think beyond the limits!

From a view of the spike glycoprotein-heparin interaction: In the PRACE research project "DyCoVin – Interactions and dynamics of SARS-CoV-2 spike-heparin complex", the Molecular and Cellular Modeling group (MCM) led by Rebecca Wade perform molecular dynamics simulations to investigate how heparin may hinder SARS-CoV-2 infection. The researchers want to characterize the structure and dynamics of putative binding patches for heparin-like compounds on the spike glycoprotein (cf. Chapter 2.8, pp. 58/59).

(Image: Giulia Paiardi).

Nahaufnahme der Interaktion zwischen Heparin und dem SARS-CoV-2 Spike-Glykoprotein: Im Rahmen des PRACE Projekts "DyCoVin – Interactions and dynamics of SARS-CoV-2 spike-heparin complex" untersucht die Gruppe „Molecular and Cellular Modeling“ (MCM) unter der Leitung von Rebecca Wade mithilfe von Molekuldynamik-Simulationen Moleküle, die am Infektionsprozess von SARS-CoV-2 beteiligt sind. Die Wissenschaftlerinnen und Wissenschaftler befassen sich insbesondere mit der Bestimmung der Struktur und Dynamik möglicher Bindungsstellen für heparin-ähnliche Wirkstoffe am Spike-Glykoprotein (siehe Kapitel 2.8, S. 58/59).  
(Bild: Giulia Paiardi).



<b>1 Think beyond the limits!</b>	<b>4</b>	<b>6 Collaborations</b>	<b>92</b>
<b>2 Research</b>	<b>8–81</b>	<b>7 Publications</b>	<b>94</b>
2.1 Astroinformatics (AIN)	8		
2.2 Computational Carbon Chemistry (CCC)	14		
2.3 Computational Molecular Evolution (CME)	18		
2.4 Computational Statistics (CST)	24		
2.5 Data Mining and Uncertainty Quantification (DMQ)	32		
2.6 Groups and Geometry (GRG)	38		
2.7 Molecular Biomechanics (MBM)	46		
2.8 Molecular and Cellular Modeling (MCM)	52		
2.9 Natural Language Processing (NLP)	60		
2.10 Physics of Stellar Objects (PSO)	66		
2.11 Scientific Databases and Visualization (SDBV)	72		
2.12 Theory and Observations of Stars (TOS)	78		
<b>3 Centralized Services</b>	<b>82</b>	<b>8 Teaching</b>	<b>98</b>
3.1 Administrative Services	82		
3.2 IT Infrastructure and Network	83		
<b>4 Communication and Outreach</b>	<b>84</b>	<b>9 Miscellaneous</b>	<b>102</b>
<b>5 Events</b>	<b>88</b>	9.1 Guest Speaker Activities	102
5.1 Conferences, Workshops & Courses	88	9.2 Presentations	104
5.1.1 Emulator Day	88	9.3 Memberships	107
5.1.2 EUStands4PM workshop “Using patient-derived data for in silico modeling in personalized medicine”	88	9.4 Contributions to the Scientific Community	109
5.1.3 ZOrA workshop	89	9.5 Awards	110
5.1.4 Workshop “FAIR Data Infrastructures for Biomedical Communities”	89		
5.1.5 Astrophysics winter workshop	90		
5.2 HITS Colloquia	90		
5.3 HITS anniversary reception	91		
		<b>10 Boards and Management</b>	<b>112</b>

# 1 Think beyond the limits!



PD Dr. Wolfgang Müller  
(Scientific Director / Institutssprecher)



Dr. Gesa Schönberger  
(Managing Director / Geschäftsführerin)

In every Annual Report, the HITS management describes the major achievements of the preceding calendar year. At the beginning of 2020, in line with previous years, we wrote a text about communication at the Institute and the role that its environment and the physical proximity of the workspaces play in fostering solidarity and joint research at HITS.

Just a few weeks later, however, came the first corona lockdown in Germany, which posed a special challenge to all of us as well as to internal communications. We were able to deal with the new requirements very thoroughly and without major setbacks because almost all scientists and non-academic staff were able to work well from home. At the same time, it was important to us to maintain the HITS spirit and remain as transparent as possible.

From the very beginning, it was critical to pass on the ever-changing public information on the implementation and easing of restrictions – which was initially only available in German – internally and on a weekly basis (and in English). As we have come to learn, the emails on the corona situation that were produced and circulated by the HITS communications team ([see Chapter 4](#)) were also passed on to colleagues at other institutes due to their thoroughness and usefulness.

Beginning in March, all events were canceled and soon switched to digital formats, which were even better attended than their pre-corona non-digital counterparts. The online end-of-the-year celebration was also a great success. We believe that our diligent communication in years past as well as during the crisis helped

keep the HITS team spirit high. This team spirit remained despite the problems that most of us had and continue to have with the pandemic. The most important part of keeping spirits high, however, is played – as always – by the HITSters themselves thanks to their interest, motivation, trust, and good will. HITS is a special place to work, even when working from home.

This positive attitude was also reflected in our research: Some HITS groups participated in corona-related research ([see Chapters 2.3, 2.4, 2.8, and 2.11](#)), while others took advantage of the time to publish a considerable number of papers. In addition, numerous third-party funding approvals were granted, including no less than three ERC grants: one each for HITS group leaders Frauke Gräter (MBM) and Saskia Hekker (TOS) as



well as for Fabian Schneider, a visiting scientist at proposal writing time who is using the funds from his ERC Starting Grant to establish his own junior group at HITS, "Stellar Evolution Theory" (SET), which began in January 2021.

We were thrilled to see independent reviewers recognize the achievements of HITS researchers in their respective fields, and we continue to work on the further development of HITS. An essential characteristic of the Institute is its interdisciplinarity. From the very moment of its founding, three clusters of research emerged and have been continually strengthened: We have groups that develop and apply methods in the life sciences, groups that make astronomical observations and simulations, and groups that work in a method-centered, cross-disciplinary way.

Within the individual fields, collaboration is relatively easy. Researchers from the life sciences, for example, share a common language. However, collaboration across disciplinary boundaries is more challenging. While there is already ongoing interdisciplinary work at HITS, we aim to give the opportunity addressing this challenge more intensively. One tool that can be used to help improve collaboration is called the "HITS Lab."

The HITS Lab is an internal funding program for projects in which at least two groups from different disciplines at HITS come together to work on a shared topic. The participating groups have the opportunity to hire researchers as part of the HITS lab who – in turn – are jointly supervised by the respective group leaders.

The first project to emerge from the initial considerations of the HITS Lab was launched toward the end of 2019, when Frauke Gräter (MBM) and Michael Strube (NLP) together with Vera Nünning (from the English Department at Heidelberg University) – working within the framework of a project at the "Marsilius Kolleg" at Heidelberg University – asked, "Does the quality of writing influence scientific impact?" Around the same time, Michael Strube (NLP), Wolfgang Müller (SDBV), and colleagues obtained funding from the BMBF for the

project "DeepCurate" ([see Chapters 2.9 and 2.11](#)). The scientific idea also came from the HITS Lab initiative. Two additional HITS Lab projects began in 2020. The first project, "Emulation in Simulation", is a collaboration between Frauke Gräter (MBM), Fritz Röpke (PSO), and Tilmann Gneiting (CST) with the aim of estimating partial results via the clever use of machine-learning techniques – so-called emulators – and thereby of reducing computational effort ([see Chapters 2.4 and 5.1.1](#)). The second project, "Geometry and Representation Learning," is a collaborative effort between Anna Wienhard (GRG), Michael Strube (NLP), and their groups that investigates the use of non-Euclidean geometries within natural language processing ([NLP, see Chapter 2.6](#)).

Despite the pandemic and contact restrictions of all kinds, there is much to report from HITS for 2020, and the next few years promise to continue to be fruitful. We expect to see projects and results that will take full advantage of the diversity of HITS and lead to the development new ideas. As you can also see, Scientific Director Frauke Gräter, who assumed office at the beginning of 2021, will continue to be heavily involved – not only in the management of HITS, but also in HITS Lab projects. We look forward to all that is to come.





In jedem Jahresbericht berichtet die Institutsleitung über das vergangene Jahr. Anfang 2020 schrieben wir bezogen auf die Erfahrungen aus den Vorjahren einen überzeugten Text über Kommunikation im Institut, sowie über die Rolle, die die Umgebung des HITS und die kurzen Wege für den Zusammenhalt und die gemeinsame Forschung am HITS spielen.

Doch schon wenige Wochen später kam der erste Corona-Lockdown in Deutschland und damit eine besondere Herausforderung für uns alle. Und auch eine Herausforderung für die interne Kommunikation. Wir sind sehr konsequent mit den neuen Notwendigkeiten umgegangen. Dies konnten wir tun, weil nahezu alle Wissenschaftler/-innen ebenso wie das nicht-wissenschaftliche Personal gut von zuhause arbeiten konnten. Zugleich war es uns ein Anliegen, den

HITS-Spirit zu wahren und möglichst transparent zu sein. Wichtig war es von Anfang an, die sich ständig ändernden und zu Beginn nur in deutscher Sprache zur Verfügung stehenden öffentlichen Informationen über Restriktionen und Wiedererlaubtes wöchentlich intern (und auf Englisch) zu kommunizieren. Die von der HITS-Kommunikation (*siehe Kapitel 4*) erstellten Rundmails zur Corona-Situation werden, wie wir hören, an Kolleg/-innen anderer Institutionen weitergereicht, weil sie so gut und nützlich sind.

Ab März wurden alle Veranstaltungen abgesagt und bald auf digitale Formate umgestellt. Diese digitalen Formate waren sogar besser besucht als die nicht-digitalen Formate vor der Corona-Zeit. Auch die online-Jahresendeier war ein großer Erfolg. Wir haben das Gefühl, dass unsere sorgsame Kommunikation in den

Jahren davor, ebenso wie die Kommunikation während der Krise, geholfen haben, den HITS-Team-Spirit auf hohem Niveau zu halten. Dies gilt trotz der Probleme, die die meisten von uns mit dieser Situation hatten und haben. Am wichtigsten hierfür sind aber immer noch die HITster selbst, ihr Interesse, ihre Motivation, ihr Vertrauen und ihr guter Wille. HITS ist ein besonderer Arbeitsplatz, auch im „Homeoffice.“

Diese positive Einstellung schlug sich auch in der Forschung nieder: Einige HITS-Gruppen haben sich an Coronamotivierten Forschungsarbeiten beteiligt (*siehe Kapitel 2.3, 2.4, 2.8 und 2.11*), andere haben die Zeit genutzt, um viel zu publizieren. Darüber hinaus gab es zahlreiche Drittmittel-Bewilligungen, unter anderem über sage und schreibe drei ERC Grants für die HITS-Gruppenleiterinnen Frauke Gräter (MBM) und

Saskia Hekker (TOS) sowie für den bisher als Gastwissenschaftler am HITS tätigen Fabian Schneider, der die Mittel aus seinem ERC Starting Grant dafür verwendet, am HITS ab 2021 seine eigene Juniorgruppe „Stellar Evolution Theory“ (SET) aufzubauen.

Die Anerkennung unabhängiger Gutachter/-innen für die Leistungen der HITS-Forscher/-innen in ihren Spezialgebieten sehen wir mit großer Freude – und arbeiten zugleich weiter an der Fortentwicklung des HITS. Ein wesentliches Merkmal des Instituts ist seine Interdisziplinarität. Schon bei der Gründung waren drei Richtungen erkennbar: Wir haben Gruppen, die Methoden in den Lebenswissenschaften entwickeln und anwenden. Wir haben Gruppen, die Astronomie beobachtend und simulierend betreiben, und wir haben Gruppen, die methodenzentriert, gebietsübergreifend arbeiten.

Innerhalb der einzelnen Gebiete ist das Zusammenarbeiten vergleichsweise einfach. Forschende aus den Lebenswissenschaften zum Beispiel haben eine gemeinsame Sprache, doch die Zusammenarbeit über Disziplingrenzen hinweg bleibt eine Herausforderung. Dieser will sich HITS in Zukunft noch intensiver stellen, ein Werkzeug hierzu heißt „HITS Lab“.

Das HITS Lab ist ein internes Förderprogramm für Projekte, in denen sich mindestens zwei Gruppen aus unterschiedlichen Disziplinen am HITS zusammenfinden, um ein gemeinsames Thema zu bearbeiten. Die beteiligten Gruppen haben die Möglichkeit, dafür Mitarbeiter/-innen einzustellen, die wiederum von zwei Gruppenleiter/-innen gemeinsam betreut werden.



Gegen Ende des Jahres 2019 lief das erste Projekt an, das aus den ersten Überlegungen zum Thema HITS Lab entstanden war: „Does the quality of writing influence scientific impact?“

fragten Frauke Gräter (MBM) und Michael Strube (NLP) gemeinsam mit Vera Nünning (Anglistisches Seminar der Universität Heidelberg) im Rahmen eines Projektes am Marsilius-Kolleg der Universität Heidelberg. Etwa zeitgleich haben Michael Strube (NLP), Wolfgang Müller (SDBV) und Mitarbeiter/-innen das Projekt „Deep-Curate“ beim BMBF eingeworben (*siehe Kapitel 2.9 und 2.11*). Die wissenschaftliche Idee kam auch hier aus der HITS Lab-Initiative.

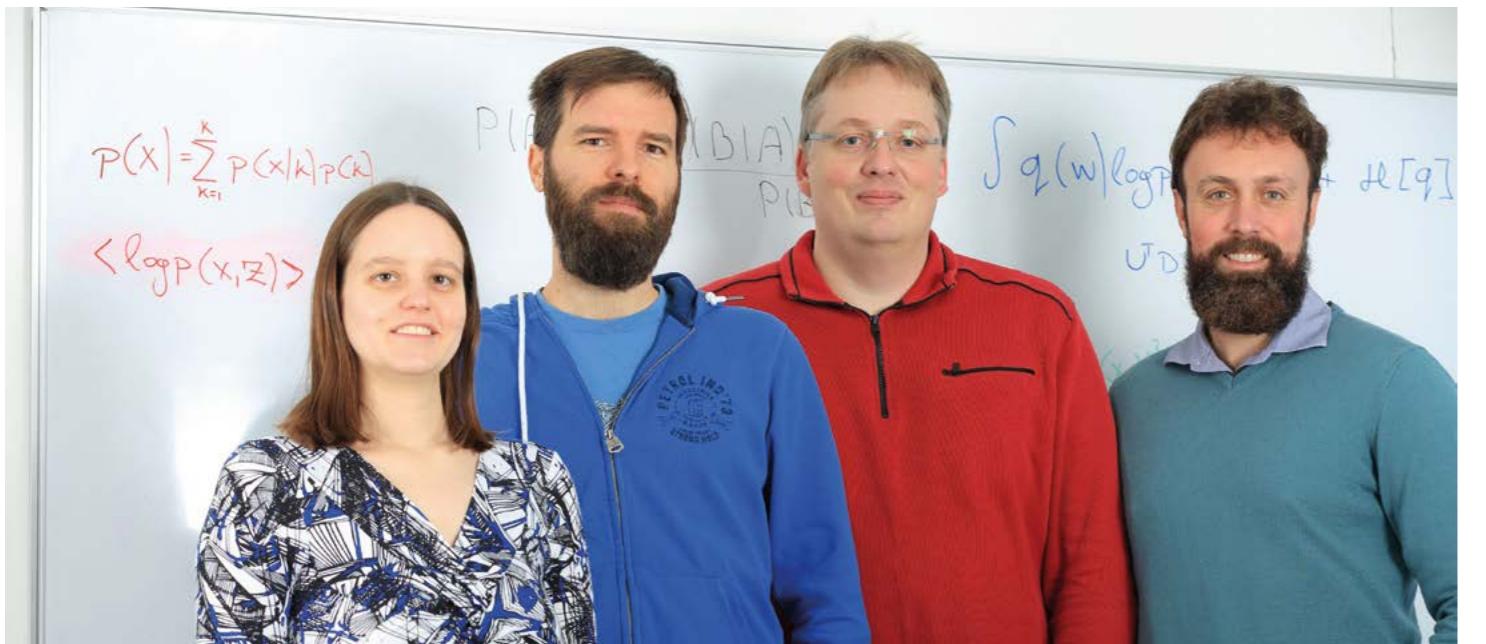
Im Jahr 2020 haben zwei weitere HITS Lab-Projekte begonnen: Zum einen das Projekt „Emulation in Simulation“, eine Zusammenarbeit von Frauke Gräter (MBM), Fritz Röpke (PSO) und Tilmann Gneiting (CST). Hier geht es darum, durch geschickten Einsatz von Techniken des maschinellen Lernens als sogenannte Emulatoren, Teilresultate zu schätzen und so den Rechenaufwand zu reduzieren (*siehe Kapitel 2.4 und 5.1.1*). Und zum anderen das Projekt „Geometry and Representation Learning“ – hier arbeiten Anna

Wienhard (GRG) und Michael Strube (NLP) mit ihren Gruppen an nicht-euklidischen Geometrien für Lernaufgaben in der natürlichsprachlichen Datenverarbeitung (*siehe Kapitel 2.6*).

Trotz Pandemie und Kontakt einschränkungen jeglicher Couleur gibt es aus dem HITS für 2020 viel zu berichten, und man sieht jetzt schon, dass die nächsten Jahre bunt werden. Wir erwarten viele interessante Projekte und Resultate, in denen wir die Vielgestaltigkeit des HITS nutzen und neue Ideen entwickeln. Wie Sie auch sehen können, ist die ab 2021 amtierende Institutssprecherin Frauke Gräter sehr stark engagiert – nicht nur in der Leitung des HITS, sondern auch bei den HITS Lab-Projekten. Wir freuen uns darauf.

## 2 Research

### 2.1 Astroinformatics (AIN)



#### Group Leader

Dr. Kai Polsterer

#### Staff members

Dr. Nikos Gianniotis (staff scientist)  
Dr. Antonio D'Isanto  
Dr. Jan Plier (since June 2020)

#### Scholarship holder

Erica Hopkins

#### Student assistant

Fenja Kollasch

In recent decades, computers have come to revolutionize 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 at higher spatial, spectral, and temporal resolutions. In addition, the possibility to observe astroparticles and gravitational waves along with previously untapped wavelength regimes is now granting more-complete access to the Universe.

The Astroinformatics group deals with the challenges of analyzing and processing such complex, heterogeneous, and large datasets. Our scientific focus in astronomy is on evolutionary processes as well as extreme physics in galaxies, as found, for example,

around active super-massive black holes in the centers of galaxies. Driven by these scientific challenges, we develop new methods and tools and share them with the community. From a computer-science perspective, we focus on time-series analyses, sparse-data problems, morphological classification, the proper evaluation and training of models, and the development of explorative-research environments. These methods and tools will prove critical to the analysis of data in large upcoming survey projects, such as SKA, Gaia, LSST, and Euclid.

Our ultimate goal is to enable scientists to analyze the ever-growing volume of information in a bias-free manner.

#### Probabilistic flux variation gradient

We have been working on a probabilistic reformulation of the flux-variation-gradient method with the goal of disentangling the roles played by active galactic nuclei (AGN) and their host galaxies in photometric reverberation mapping. This work will serve as a precursor study before we embark on developing more-powerful models that consider the actual physics that underlie the generation of the observed light curves. Such models will help us not only in disentangling the photometric contributions of AGNs and their host galaxies but also in shedding light on the physical properties of these systems, such as their black-hole mass and accretion rate.

AGNs contain large black holes in their centers and produce a great deal of energy, which renders them among the most-luminous objects in the Universe. Due to their extremely compact size with respect to their host galaxies and to their usually

methods have relied on fitting galaxy templates and modeling the host-galaxy profile via high-resolution images. The flux-variation-gradient (FVG) method does not require images of high spatial resolution and instead simply makes use of fluxes measured at different photometric bands. However, FVG does require (i) a constant-in-time contribution of the host galaxy (including non-varying emission lines), (ii) a varying AGN contribution, (iii) an empirically derived linear relationship of fluxes in different photometric bands, and (iv) knowledge of the colors of the host galaxy in question.

The FVG can be best understood geometrically, as explained in Figure 1. In this simple geometric view, the goal of FVG is to find the intersection point between the line of the vector (termed the "galaxy vector"; green in Fig. 1) that expresses the hypothesized colors of the host galaxy and the line (dashed in Fig. 1) on which the total flux (galaxy plus AGN) observations lie (pink in Fig. 1). Inferring this intersection point

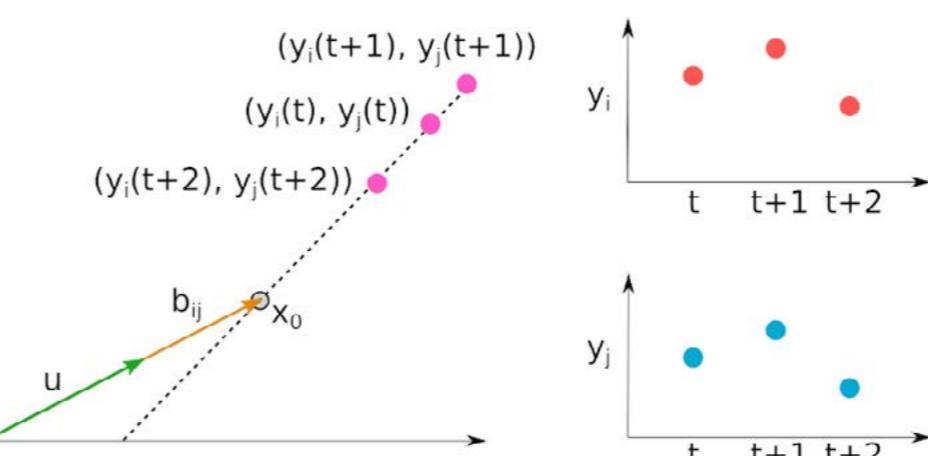


Figure 1: Sketch of FVG: On the right-hand side, we see observations from bands  $i$  (top) and  $j$  (bottom) measured at three different time instances. By pairing flux values of co-occurring observations, we form points (in pink) in the flux-flux plot (left-hand side) that fall on a dashed line. Vector  $b_{ij}$  (in brown) corresponds to the unobserved host galaxy and defines line  $b_{ij} \cdot x$ , which intersects the dashed line at  $x_0$ . The FVG method consists of finding the intersection of  $u \cdot x$  and the dashed line, where  $u$  (in green) is the so-called galaxy vector that is assumed to have the same direction (i.e., the same colors) as the unobserved  $b_{ij}$ .

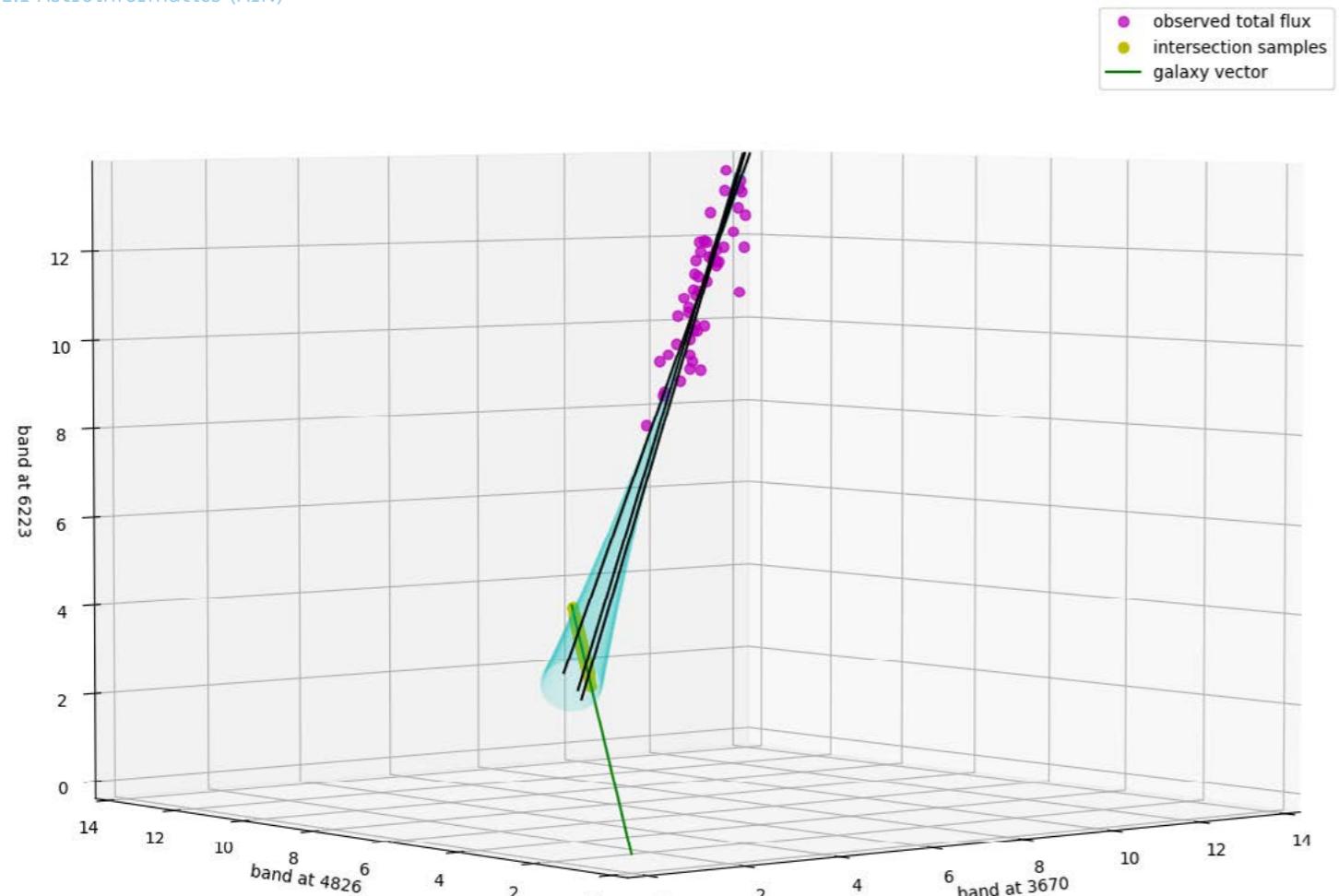
relatively large distance from Earth, AGNs appear as point sources and are difficult to spatially resolve in photometric observations. Previous

the FVG on a probabilistic basis in order to endow it with the ability to (i) account for uncertainty in the flux measurements, (ii) jointly take all photometric bands into account when inferring the intersection point, and (iii) produce a distribution – as opposed to a point estimate – of where the intersection point is likely to be located.

Our probabilistic reformulation comprises two steps. The first step involves identifying the total flux line (dashed line in Fig. 1) as the first principal component obtained via probabilistic principal component analysis (PPCA). In contrast to classical principal component analysis, PPCA incorporates an explicit noise model that allows us to account for the presence of noise in the observed data. Furthermore, by adopting a Bayesian perspective, we can work out a distribution of likely first principal components, which means that we can identify a set of likely lines on which the observed fluxes may lie. This is illustrated in Figure 2 for the case of three observed band filters.

The second step of our approach involves identifying the intersection point. The current FVG method identifies the intersection point as the intersection of a single line that passes through the total flux observations (pink points in Fig. 2) and the line implied by the galaxy vector (e.g.,  $u$  in Fig. 2; in green). In our approach, as illustrated in Figure 2, we have a distribution of lines as opposed to a single line. Hence, we need to clarify what it means to search for an intersection point between a line (defined by the galaxy vector) and a density of lines: The intersection we seek is a point along the line implied by the galaxy vector (green in PF2) that receives considerable support by the density of lines. This view leads to a distri-

## 2.1 Astroinformatics (AIN)



**Figure 2:** Probabilistic FVG in action: The pink dots represent observations in three photometric bands; the green line corresponds to the so-called galaxy vector, which is hypothesized to have the same colors as the host galaxy. Our method infers a distribution of lines that likely go through the observations (pink dots). The cyan tube contains the distribution of these lines. We plot three samples of such lines in black and note that the uncertainty of the tube grows with increasing distance from the observed data. The yellow dots display the distribution of possible intersection points – that is, the “intersection” of the green line with the density of lines that go through the observed data.

bution of likely intersection points (yellow dots in Fig. 2) and hence to a distribution of likely relative photometric contributions by the AGN vs. the host galaxy.

### Infrastructure for exploring astronomical spectra

Data archives have a long history of use in astronomy. In the past, the main preservation media were photographic plates and catalogs. Since the beginning of the digital era, however, digital data archival has been a major topic in astronomy. Advances in instrumentation and dedicated data-intense survey telescopes have led to an increasing demand for novel data infrastructures that make efficient data

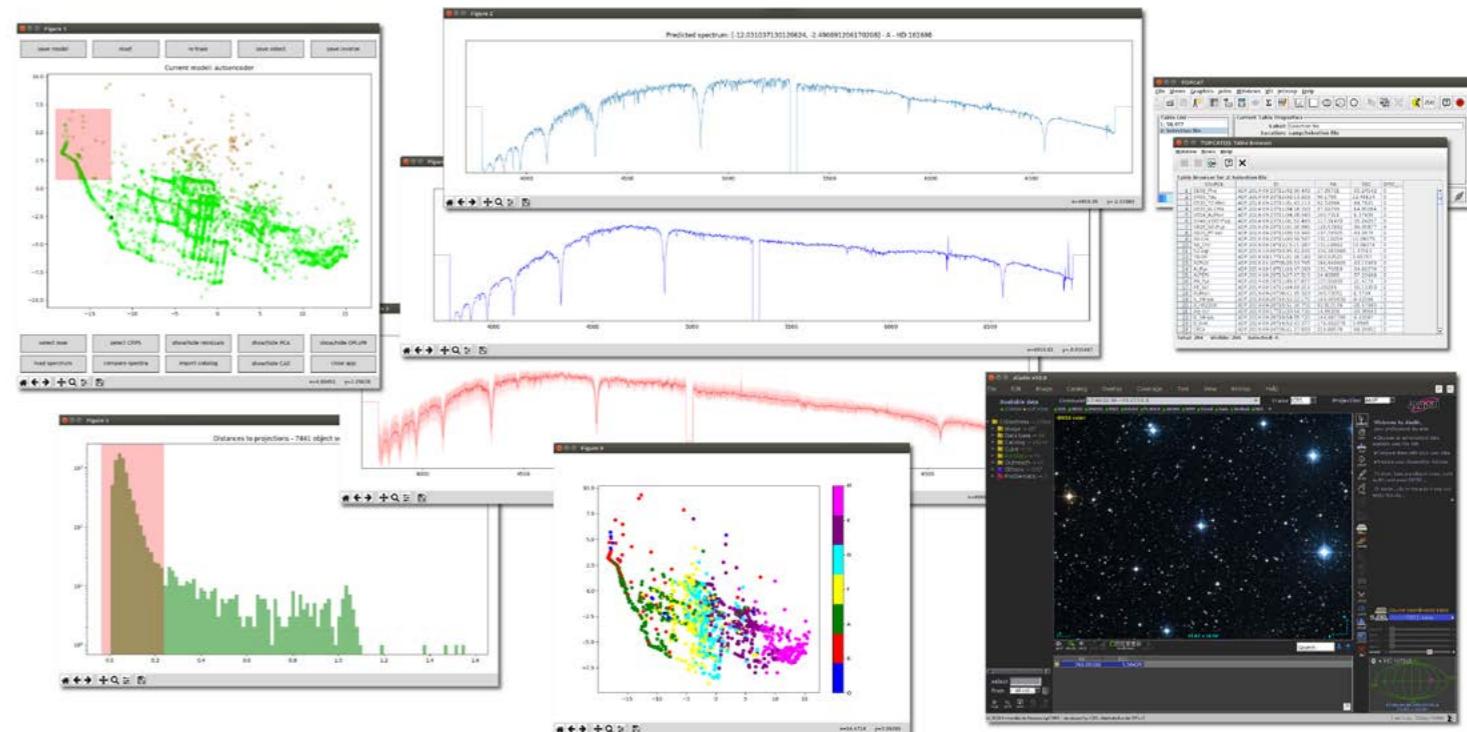
access and analysis possible. The astronomical community has made a large effort to fulfill these needs and provide the required infrastructure. In so doing, the role of the International Virtual Observatory Alliance (IVOA) has been essential as the IVOA defines standards and ensures a proper discussion on how to make infrastructures, software, and data interoperable and to how follow concepts like the FAIR data principles. However, the current services of data providers reveal severe limitations in functionality in the context of the Big Data regime. The data deluge has rendered traditional access- and analysis techniques unfeasible with respect to the size of modern archives.

Modern practice has come to adopt machine-learning solutions to deal with a wide range of analysis tasks used in answering astronomical questions. However, these solutions require efficient access to and processing of a vast number of data in order to train models and obtain reliable results. Some data centers are currently working on concepts such as “bringing code to the data” in order to overcome bottlenecks related to data transfer. Other approaches focus, for example, on data visualization, predefined analysis tasks, or queries about pre-extracted features that provide a compressed representation of the original data.

In this context, the AIN group at HITS joined the ESCAPE project, a large European collaboration whose aim is to tackle the new challenges created by data-driven research in astronomy and astroparticle physics while maintaining a joint focus on dealing with complex data workflows, infrastructural issues, and data- and software interoperability. Our contribution was the develop-

from the field of machine learning combined with a simple graphical user interface that allows for compressed visualization, inspection, and interaction with spectral observations. Dimensionality-reduction methods are used to construct a low-dimensional space into which high-dimensional data items are projected with the aim of projecting similar objects close to one another. This method enables us

an archive that contains spectral information from heterogeneous sources. This prototype is meant to demonstrate how we can overcome the need to download a complete dataset in order to find a few interesting sources. This new concept of data access and interaction could become an additional standard for accessing the next generation of archives (see Figure 3).



**Figure 3:** The prototype developed to explore the spectra in the HARPS archive. The overview panel allows for interaction with the ML models and for selecting data (upper left). The characteristics of the spectra at the selected coordinate and in the selected area are shown in the three corresponding spectral plots (center). Histogram functions for subset selection as well as an overplot of spectral classes are shown at the bottom. VO tools – such as Topcat (for inspecting the selected sources as tables) and Aladin (for inspecting the corresponding images) – are integrated through VO standards (right).

ment of new explorative access methods for spectroscopic data taken from the ESO archive. Instead of following current data-retrieval practices, we used explicit search criteria and positionally indexed queries to come up with a novel search paradigm based on structural information as well as on data similarity. In other words, we built a prototype that allows implicit and explorative access to data, including searches by similarity.

We reached our goal by utilizing dimensionality-reduction methods

to browse massive datasets ordered by structural similarities in order to find classes, outliers/anomalies, and scientifically relevant objects. For the prototype, we chose an autoencoder as the main dimensionality-reduction model: an unsupervised neural network that is able to compress original data into a low-dimensional representation and to generate a reconstruction from this low-dimensional space.

Two datasets were selected as use cases: HARPS, an archive that includes stellar spectra only, and UVES,

In addition to multiple similarity measures, several dimensionality-reduction models beyond a plain autoencoder were implemented (Principal Component Analysis, Gaussian Process Latent Variable Model, variational autoencoder, convolutional autoencoder). Furthermore, the prototype is connected to the most-important VO tools (Aladin, Topcat, Splat) in order to obtain further information on the sources and allow additional interactive features. The user can explore the projected space by similarities, make selections, create new catalogs, and

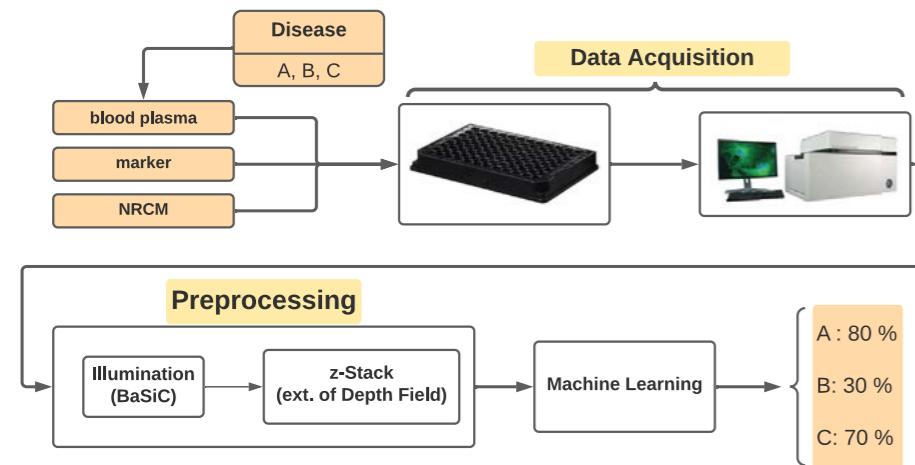


Figure 4: The long-term goal of the project presented as a flow chart. The current stage uses special substances as surrogates for the diseases. The pipeline – including the pre-processing phase – has already been implemented.

import and export data. It is possible to load pre-trained models, to retrain them, and to select between different loss functions.

Inspecting the projections of the HARPS dataset yielded a clear sequence of structural similarities that corresponds to the main stellar spectral classes, which was confirmed by checking the spectral classification from Simbad and over-plotting it as color. This result was expected because the autoencoder had been configured to project the spectra in a 2-D space and to allow the projections to be visualized

In den letzten Jahrzehnten hat der Einsatz von Computern die Astronomie stark beeinflusst. Der technologische Fortschritt ermöglichte den Bau neuer Detektoren und innovativer Instrumente sowie neuartiger Teleskope. Damit können Astronomen nun mehr Objekte als je zuvor mit bisher unerreichtem Detailreichtum, sowohl räumlich, spektral als auch zeitlich aufgelöst beobachten. Hinzu kommen neue Beobachtungsmöglichkeiten durch z.B. Astroteilchen sowie Gravitationswellen, die neben bisher nicht beobachtbaren Wellenlängenbereichen ein vollständigeres Bild des Universums bieten. Die Forschungsgruppe **Astroinformatik (AIN)** beschäftigt sich mit den Herausforderungen, die durch die Analyse und Verarbeitung dieser komplexen, heterogenen und großen Daten entstehen. In der Astronomie beschäftigen uns die Fragestellungen im Bereich der Galaxienentwicklung sowie die extremen physikalischen Vorgänge, wie man sie z.B. in der Umgebung von aktiven supermassereichen schwarzen Löchern in den Zentren von Galaxien findet. Auf diesen Fragestellungen basierend, entwickeln wir neue Methoden und Werkzeuge, die wir frei zur Verfügung stellen. In der Informatik liegt unser Interesse hierbei auf der Zeitreihenanalyse, dem Umgang mit spärlichen Daten, der morphologischen Klassifikation, der richtigen Auswertung und dem richtigen Training von Modellen sowie explorativen Forschungsumgebungen. Diese Werkzeuge und Methoden sind eminent wichtig für aktuelle und sich gerade in der Vorbereitung befindenden Projekten, wie SKA, Gaia, LSST und Euclid.

Unser Ziel ist es, einen möglichst unvoreingenommenen Zugang zu dieser enormen Menge an Information zu gewährleisten.

Computational cardiology

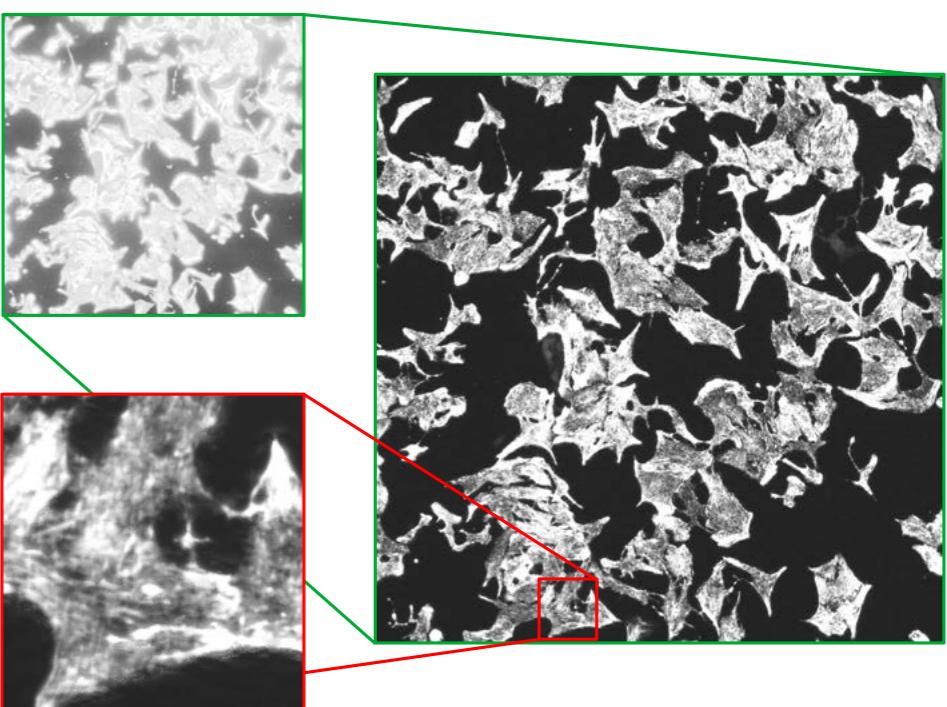
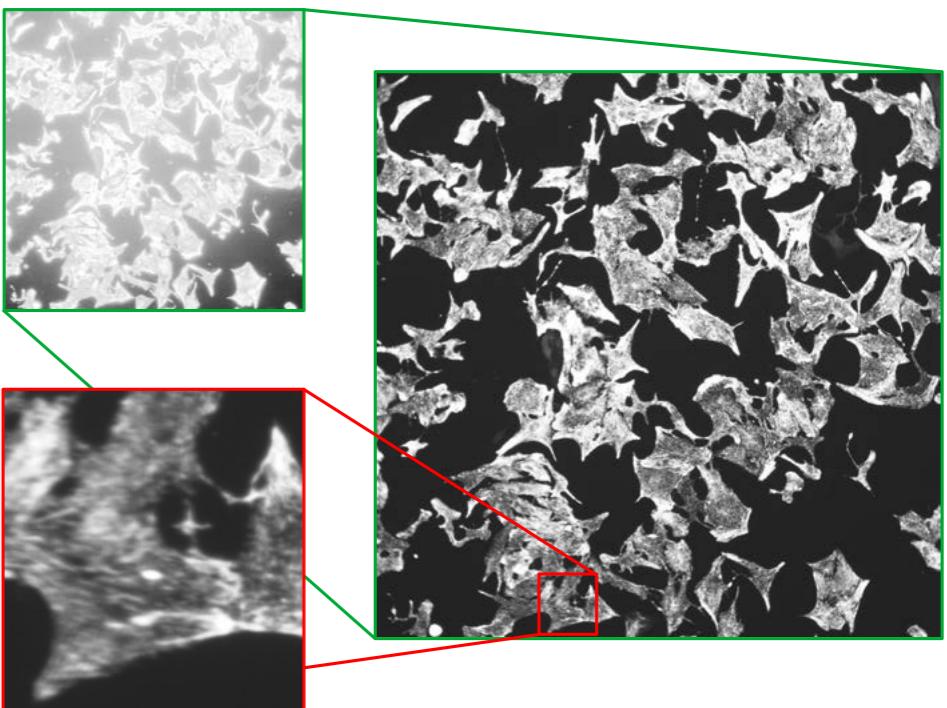
Due to our interest in analyzing massive datasets as well as in morphologically analyzing galaxies at large scales, we are also involved in a medical project dealing with morphological data at microscopic scales (see Chapter 6, "Informatics4Life"). Our role in this project involves building a pipeline that processes the captured images and extracts features that describe the morphological structure of each captured cell. Subsequently, these extracted

on a screen. With such a configuration, the model mainly learns to reconstruct the spectral continuum, which is directly connected to temperature and therefore also to the spectral classes. features will be used to build a robust and stable diagnostic tool. This project does not deal with data from astronomy, but rather with image data from a microscope instead of a telescope. Both con-

Our future plans include the possibility of transferring the prototype to a Jupyter Notebook/Lab package, challenges, and it is therefore worth transferring solutions from one field to the other.

Instead of extracting morphological features from galaxies, we are using our profound knowledge in this area to develop a machine-learning approach to detect single cells and describe their morphological structures.

ture individually at the single-cell level. The ability to extract interpretable features will enable us to determine features relating to a specific substance, which is necessary to ensure reliable application in cardiology. We therefore use selected engineered features, automatically learned features, and compressed representations as the basis for a balanced comparison, for example, between pure classification performance and scientific robustness. First experiments with respect to unsupervised machine learning are currently underway (for the long-term goal of the project, see Fig. 4).



*Figure 5: Comparison of the image quality gained by applying different pre-processing steps (before: above; after: below). Note the improvement in background illumination (red) and in sharpness (green).*

## 2 Research

### 2.2 Computational Carbon Chemistry (CCC)



#### Group Leader

Dr. Ganna Gryn'ova

#### Staff members

Dr. Christopher Ehlert  
Anna Piras

#### Scholarship holder

Oğuzhan Kucur

#### Visiting scientist

Dr. Michelle Ernst (SNSF Scholarship,  
since August 2020)

#### Project student

Juliette Schleicher (Heidelberg University,  
May–July 2020)



Modern functional materials combine structural complexity with targeted performance and are utilized across many areas of industry and research, from nanoelectronics to large-scale production. Theoretical studies of these materials bring mechanistic underpinnings to light, facilitate the design and pre-screening of candidate architectures, and ultimately enable predictions of the physical and chemical properties of new systems to be made.

The Computational Carbon Chemistry (CCC) group uses theoretical and computational chemistry to explore and exploit diverse functional organic and hybrid materials. In its 2nd year at HITS, the group focused on establishing reliable yet computationally feasible protocols for the structural exploration and property quantification of various systems. Diverse methods of density functional theory and wavefunction theory were tested in simulations in which graphene

and its derivatives interacted with small molecules via physi- and chemisorption and in simulations in which metal-organic frameworks encapsulated small molecules in their pores. Accurate approaches to energy and property computations were identified via benchmarking against available experimental and high-level *in-silico* data. The utility of various tools for visualizing and analyzing interactions in these complex systems was validated.

In terms of applications, the design guidelines for graphene-based sensors for nitroaromatic pollutants and for nanographene electrocatalysts for oxygen reduction reaction were elucidated, with candidate architectures being pre-screened using the established computational protocols. These findings lay the groundwork for the *in silico* development of new and improved functional carbonaceous materials and hybrid organic–inorganic materials with targeted sensing, catalytic, and electronic behavior.

#### Accurate models of gas adsorption on graphene

Christopher Ehlert, Anna Piras, and Ganna Gryn'ova

Graphene is a two-dimensional sheet of  $sp^2$  carbon atoms arranged in a honeycomb lattice. This material displays a number of tantalizing electronic and optical properties pertinent to its applications in nanoelectronic devices. Among these devices, graphene-based gas sensors are perhaps the most well-developed and utilized in practice. These sensors rely on a change in electric properties triggered by (non-)covalent

insights, they are challenged by the periodic nature of the graphene substrates and the broad variability of the adsorption complex geometries.

In order to develop a reliable protocol for simulating graphene-based gas sensors, we benchmarked a range of theoretical procedures against available experimental data for the adsorption of carbon dioxide ( $CO_2$ ) on a pristine graphene surface. Simulations across a range of models – from infinite periodic graphene to its finite clusters of varying sizes – that capture diverse  $CO_2$  adsorption geometries and employ methods based on density functional theory

with experimental results from the literature for the  $CO_2$ -graphene system. These results represent an important milestone for the CCC group and form the basis of the methodological approaches used in our ongoing and future studies on graphene chemistry. Even more strikingly, certain properties of the investigated adsorbates were found to be independent of the model- and method choice. Specifically, the relative stabilities between the different  $CO_2$  adsorption sites on pristine graphene were found to be generally independent of the density functional and the cluster size. Moreover, to arrive at an accurate

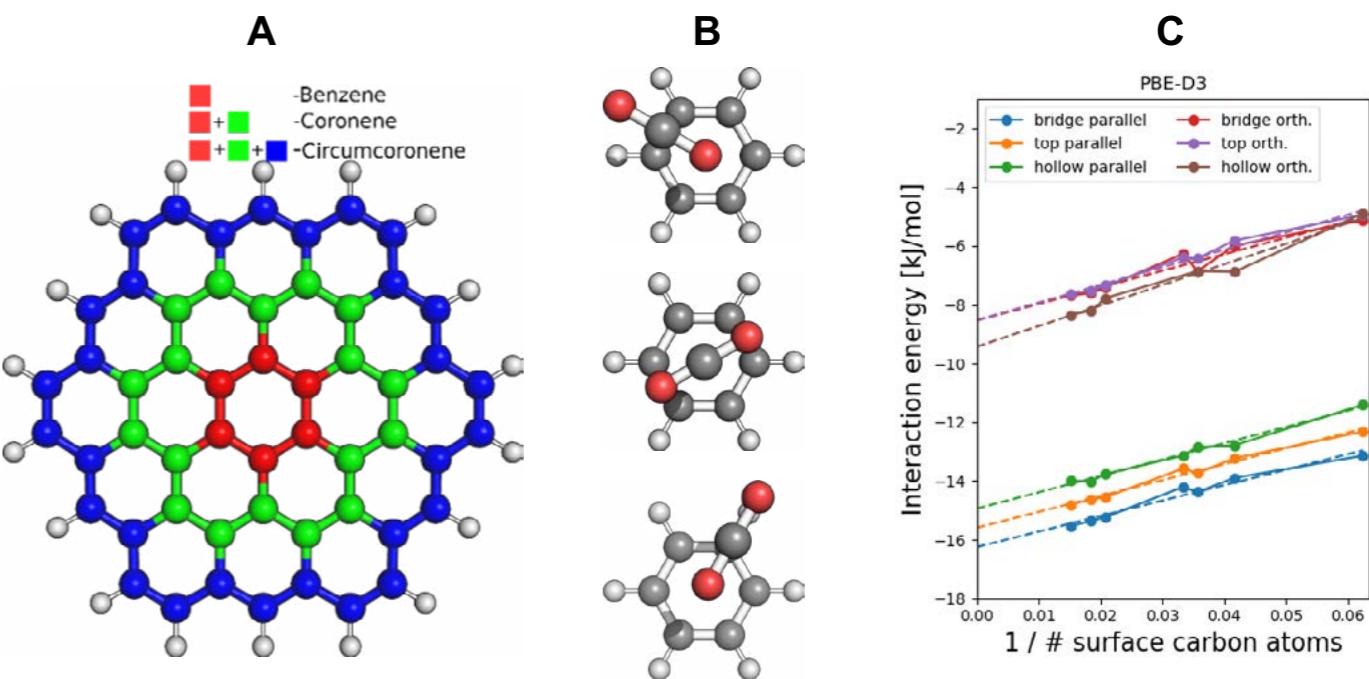


Figure 6: **A** Circular models of pristine graphene (the color code reflects the cluster size: smallest in red, medium in red and green, and largest in red, green, and blue). **B** Various orientations (bridge, hollow, and top) of  $CO_2$  adsorbed on benzene. **C** The interaction energies of  $CO_2$  with graphene as a function of the inverse number of carbon atoms in the underlying surface model at the PBE-D3 level of theory, as well as the linear regression of these data points.

interactions between the gas molecule adsorbate and the graphene-based sensing surface. A profound understanding of adsorbate-surface interactions is crucial for improving existing sensors and developing new ones with targeted selectivities and sensitivities. While accurate *in silico* simulations are indispensable for gaining the necessary mechanistic

and wavefunction theory were performed (Figure 6A, B). Through these simulations, we identified theoretical procedures for obtaining geometries and interaction energies accurately and at an acceptable computational cost compared with the prohibitively expensive gold standard of computational chemistry (the coupled cluster method) and

adsorption energy for a realistic periodic graphene sheet, we established a simple linear fit. With such an extrapolation, interaction energies for an artificial, infinitely large graphene model could be obtained within 1  $kJ\ mol^{-1}$  accuracy using any quantum-chemical method that is applicable to finite cluster models (Figure 6C).

## Nanographene catalysts for oxygen reduction reaction

**Christopher Ehlert, Anna Piras, Juliette Schleicher, and Ganna Gryn'ova**

Oxygen reduction reaction (ORR) lies at the heart of sustainable energy-conversion technologies, such as fuel cells and metal-air batteries. However, its slow kinetics necessitates the use of electrocatalysts. Recently, metal-free heteroatom-doped carbon catalysts have emerged as more efficient, stable, low-cost, earth-abundant alternatives to conventional Pt-based systems and have demonstrated excellent performance in ORR. However, the chemical complexity of such periodic systems precludes both precise mechanistic analysis and an elucidation of the structure–activity relationships. This challenge is eradicated in smaller derivatives (nanographenes, also called graphene nanoflakes or graphene quantum

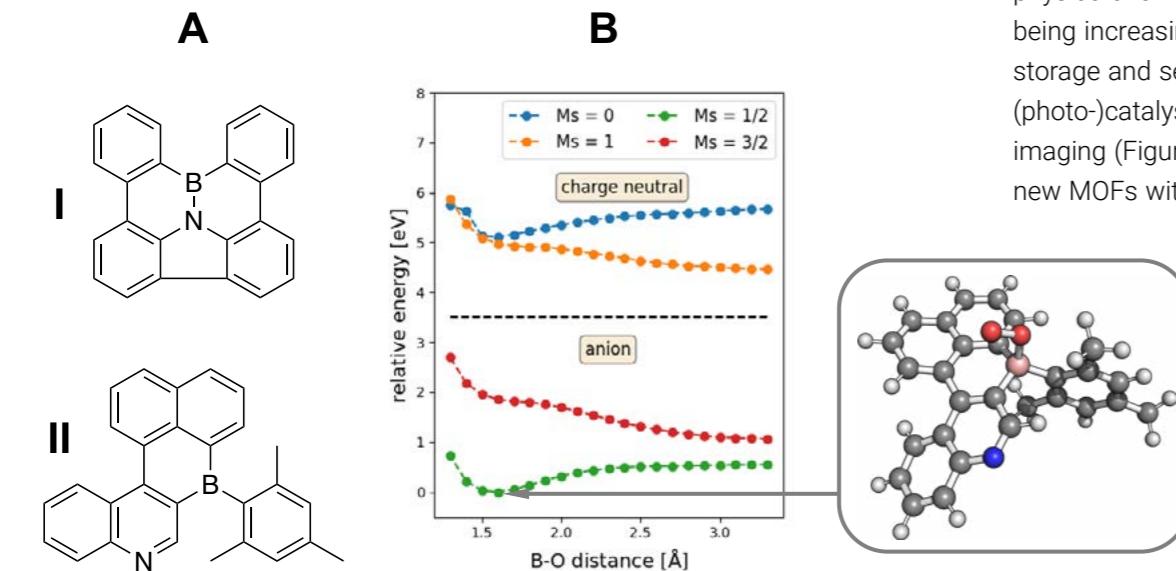


Figure 7: **A** Investigated nanographene electrocatalysts for the ORR. **B** Computed potential energy surfaces of O<sub>2</sub> chemisorption on catalyst II and the structure of the energy minimum on the MS = ½ surface (PBE-D3 level of theory).

dots), which display superior catalytic performance. To identify the physical underpinnings of this performance, we focused on two N,B-codoped nanographene

catalysts for ORR that had previously been tested experimentally (I and II in Figure 7A). The chemisorption of molecular oxygen on these substrates, which is the first step in the catalytic cycle, was simulated in several electronic states using dispersion-corrected density functional theory. First, results obtained to date highlight the important theoretical implications for modeling the ORR with electrocatalysts and suggest that electron affinity is as a simple yet powerful descriptor of the catalytic activity of nanographenes in ORR.

the electron affinities (EAs) of the two catalysts supports these findings: The computed EA of system II was higher than that of system I, which strengthened our argument on the anionic nature of the active catalytic state. These results highlight the important theoretical implications for modeling the ORR with electrocatalysts and suggest that electron affinity is as a simple yet powerful descriptor of the catalytic activity of nanographenes in ORR.

## Tools for analyzing and visualizing non-covalent interactions in metal-organic frameworks

**Michelle Ernst and Ganna Gryn'ova**

Metal-organic frameworks (MOFs) are porous crystalline hybrid organic-inorganic materials that consist of regularly connected nodes and linkers, have high internal surface areas and low densities, and are able to host small guest molecules. Due to their highly tunable composition, topologies, and physico-chemical properties, MOFs are being increasingly utilized for gas storage and separation, drug delivery, (photo-)catalysis, and biological imaging (Figure 8). A rational design of new MOFs with targeted absorption

properties requires an in-depth theoretical understanding of their microscopic building blocks and the interactions between these building blocks.

To address this challenge, we tested the applicability and validity of various quantum-chemical tools for analyzing and visualizing the strength and physical nature of the non-covalent interactions in the MOF host–guest complexes across periodic and finite-size scales. Several methods

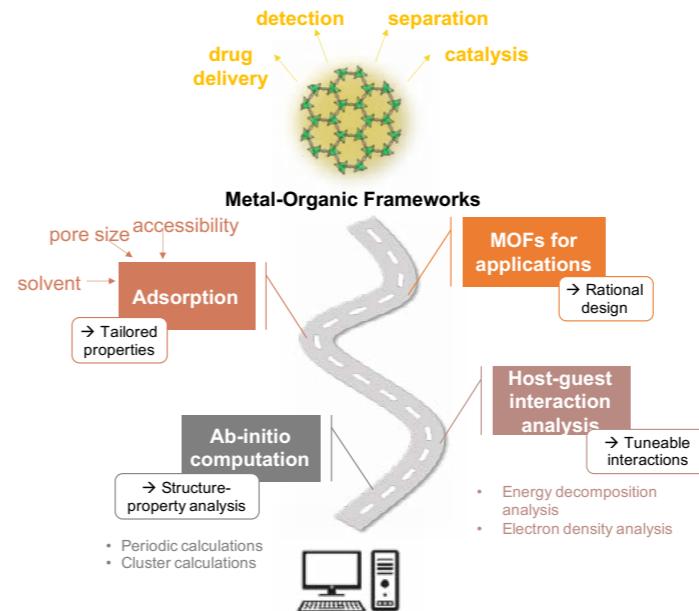


Figure 8: A roadmap for experimentally and computationally exploring MOFs.

of density functional theory in addition to symmetry-adapted perturbation theory were assessed for computing the interaction energies of the complexes of a selected MOF with two biologically active molecules in conjunction with periodic and finite cluster models (Figure 9A). Furthermore, various energy decomposition schemes were employed to elucidate the physical nature of these interactions. Next, a number of electron density partitioning tools, including the

quantum theory of atoms in molecules (QTAIM), the non-covalent interactions (NCI) index, and the density overlap regions indicator (DORI), were applied to enable visualization and further analysis of the interactions between the framework and the guest molecules (Figure 9B). Our results emphasize the critical role of model choice and corroborate the understanding that, for an appropriately chosen model, the various tested tools provide invaluable insights into the physical nature, spatial localization, and energetic strength of the interactions between MOFs and small molecules inside their pores. These findings lay the groundwork for the design and pre-screening of future MOFs without relying on pre-existing – and currently rather rare – high-resolution experimental structures.

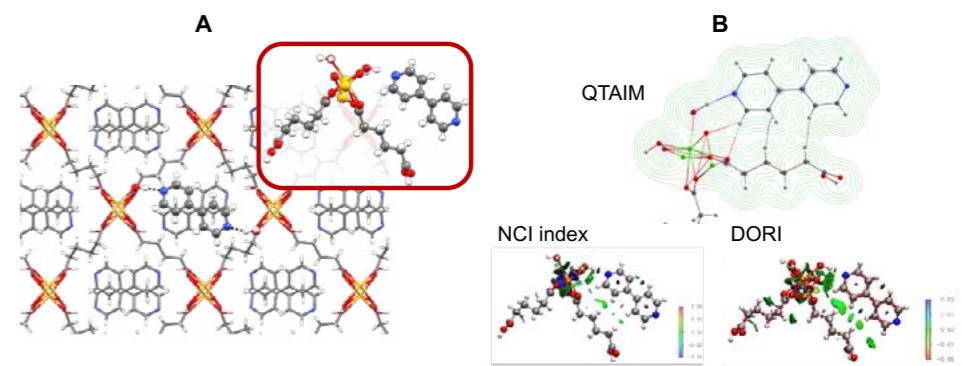


Figure 9: **A** Crystal packing of the studied MOF with a 4,4'-bipyridine guest molecule, view in c direction. One guest molecule and the nearest water molecule within the framework are shown with balls and sticks, while the corresponding hydrogen bonds between them are denoted with dashed lines. Inlet: the chosen cluster model for each complex. **B** Visualization of the results from various schemes for density partitioning.

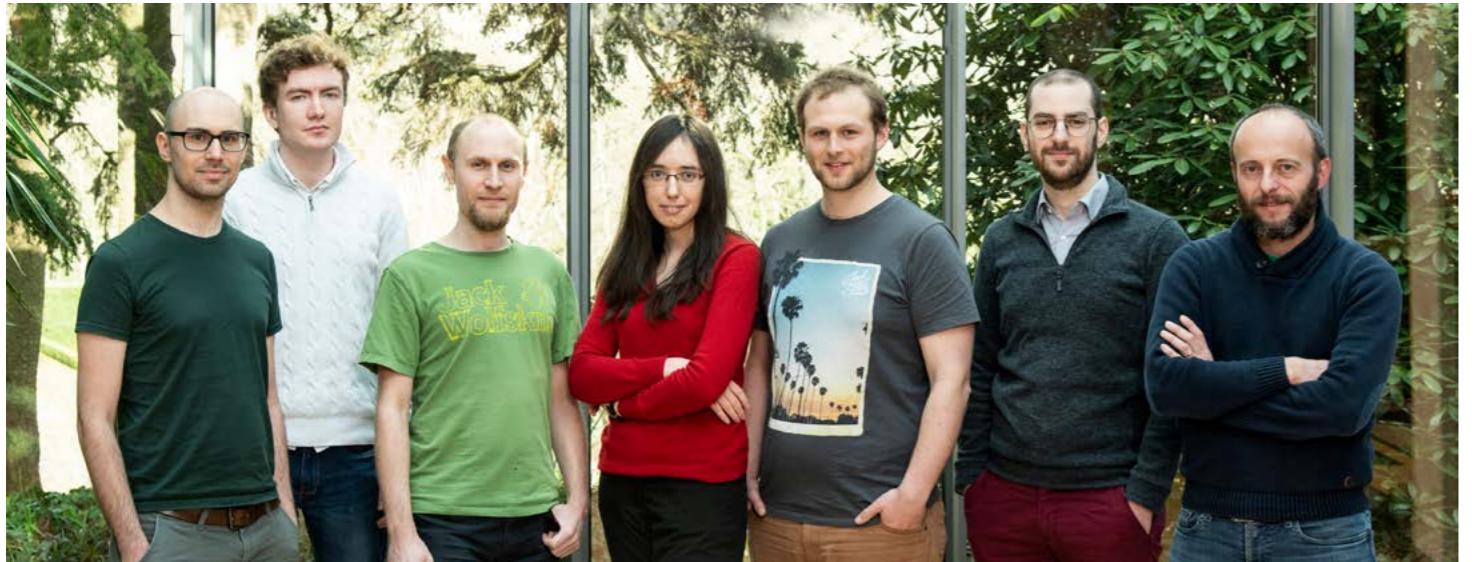
Moderne funktionale Materialien kombinieren strukturelle Komplexität mit zielgerichteter Performance und werden in verschiedenen Bereichen von Industrie und Forschung verwendet, von der Nanoelektronik bis hin zur Massenfertigung. Theoretische Studien dieser Materialien fördern mechanistische Grundlagen zutage, erleichtern das Design und Vorsortieren von Kandidaten und ermöglichen Vorhersagen zu physikalischen und chemischen Eigenschaften neu geschaffener Systeme.

Die Forschungsgruppe **Computational Carbon Chemistry (CCC)** arbeitet mit den neuesten Methoden der theoretischen und computergestützten Chemie, um funktionale organische und Hybrid-Materialien zu untersuchen und auszuwerten. In ihrem zweiten Jahr am HITS lag der Forschungsschwerpunkt auf der Erstellung zuverlässiger und gleichzeitig rechnerisch durchführbarer Protokolle für die strukturelle Untersuchung und die Quantifizierung von Eigenschaften verschiedener Systeme. Unterschiedliche Methoden der Dichtefunktionaltheorie und Wellenfunktionstheorie wurden in Simulationen getestet, bei denen Graphen und seine Derivate mit kleinen Molekülen via Physi- und Chemisorption interagierten, sowie in Simulationen, in denen metallorganische Gerüstverbindungen kleine Moleküle in ihren Poren einschließen. Hochgenaue Ansätze zur Berechnung von Energie und Eigenschaften wurden mittels Benchmarking gegen verfügbare experimentelle und high-level In-Silico-Daten bestimmt. Die Nützlichkeit verschiedener Tools zur Visualisierung und Analyse von Interaktionen in diesen komplexen Systemen wurde validiert.

Im Bereich Anwendungen wurden die Designrichtlinien für graphenbasierte Sensoren für nitroaromatische Gefahrstoffe und für nanographenbasierte Elektrokatalysatoren für die Sauerstoffreduktionsreaktion untersucht. Dabei wurden Kandidaten anhand etablierter computergestützter Protokolle vorsortiert. Die Forschungsergebnisse bilden die Grundlage für die In-Silico-Entwicklung neuer und besserer funktionaler kohlenstoffhaltiger und hybrider organisch-anorganischer Materialien mit gezielten sensorischen, katalytischen und elektronischen Eigenschaften.

# 2 Research

## 2.3 Computational Molecular Evolution (CME)



### Group Leader

Prof. Dr. Alexandros Stamatakis

### Staff members

Dr. Alexey Kozlov (staff scientist)  
Benjamin Bettsworth  
Benoit Morel  
Lukas Hübner (from October 2020)  
Pierre Barbera  
Sarah Lutteropp

### Students

Dimitri Höhler  
Ivo Baar (until April 2020)  
Johanna Wegmann (until March 2020)  
Julia Schmid (as of December 2020)  
Paul Schade (until November 2020)  
Lukas Hübner (until June 2020)

The Computational Molecular Evolution group focuses on developing algorithms, models, and high-performance computing solutions for bioinformatics. We focus mainly on

- computational molecular phylogenetics
- large-scale evolutionary biological data analyses
- supercomputing
- quantifying biodiversity
- next-generation sequence data analyses
- scientific software quality & verification

Secondary research interests include

- emerging parallel architectures
- discrete algorithms on trees
- population genetics

In the following section, we outline our current research activities, which lie at the interface(s) between computer science, biology, and bioinformatics. The overall goal of the group 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, we strive 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 that can be analyzed, observed, and tracked at the DNA level. As evolutionary biologist Theodosius Dobzhansky's famous and widely quoted dictum states, "Nothing in biology makes sense except in the light of evolution."

### What happened in the lab in 2020?

In the winter of 2019/2020, Alexis, Ben, Alexey, and Pierre taught the "Introduction to Bioinformatics for Computer Scientists" class at the Karlsruhe Institute of Technology (KIT). As in previous years, we received highly positive teaching evaluations from the students (with a learning quality index of 100 out of 100; see [http://cme.h-its.org/exelixis/web/teaching/courseEvaluations/Winter19\\_20.pdf](http://cme.h-its.org/exelixis/web/teaching/courseEvaluations/Winter19_20.pdf)).

During the summer semester of 2020, we again taught our main seminar, "Hot Topics in Bioinformatics." Our teaching activities were heavily affected by the current pandemic. The seminar in the summer was carried out entirely online. In addition, the vast majority of the oral exams for the class from winter 19/20 were also conducted online. In general, the transition to pure online teaching was comparatively easy and unproblematic. Nonetheless, students at KIT appear to now be tired of the online teaching and miss having social contact with others on the university campus.

Lukas Hübner, Dimitri Höhler, and Paul Schade all successfully defended their master's theses at the Department of Computer Science at KIT. The supervision of their theses was also conducted online to a very large extent.

We are happy that Lukas Hübner joined the lab as a PhD student in October 2020. He is co-supervised by and co-financed along with Prof. Peter Sanders, head of the algorithm engineering group at the Institute for Theoretical Informatics at KIT.

In 2020, a total of four KIT master's students joined the lab either as student programmers – to work on their master's theses – or as PhD students.

Our recurring highlight, the summer school on Computational Molecular Evolution on Crete, for which Alexis again served as main organizer in the 12th year and which was scheduled to take place in May 2020, was postponed until 2021 due to the pandemic. Moreover, the 13th iteration of the summer school, which was scheduled to take place in Hinxton, UK, in 2021, was postponed until 2022. Overall, the crisis management worked well as the decision to postpone the event was made sufficiently early.

Alexis was listed on the Clarivate Analytics list of highly cited researchers for the fifth year in a row as well as for the third consecutive year under the new "cross-field" category, which comprises researchers with a focus on interdisciplinary research (see Chapter 9.5).

The year was dominated by the pandemic. Overall, the transition to working online was relatively easy as online conferencing and collaboration tools – such as slack, github, overleaf, etc. – had already been in use long before the pandemic began, and lab members were already acquainted with online supervision. We also established a Friday coffee break conference call to maintain some form of social life at the lab.

### Introduction

The term "computational molecular evolution" refers to computer-based methods of reconstructing evolutionary trees from DNA or – for example – from protein- or morphological data. The term also refers to the design of programs that estimate statistical properties of populations – that is, programs that disentangle 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. Today, in the age of the molecular data avalanche, the manual reconstruction of trees is no longer feasible. Evolutionary biologists thus have to rely on computers and algorithms for phylogenetic and population-genetic analyses.

Following the introduction of so-called short-read sequencing machines (machines used by biologists in the wet lab to extract DNA data from organisms), which can generate over 10,000,000 short DNA fragments (each containing between 30 and 400 DNA characters), the community as a whole now faces novel challenges. One key problem that needs to be addressed is the fact that the number of molecular data available in public databases is growing at a significantly

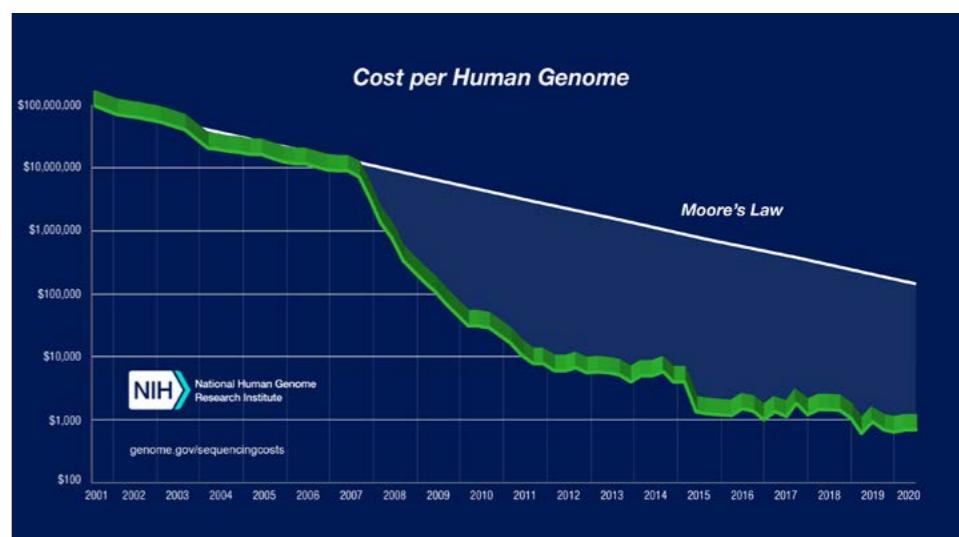


Figure 10: Cost of sequencing a human genome over time in comparison with the cost of computing according to Moore's law (source: National Human Genome Research Institute).

faster rate than the computers that are capable of analyzing the data can keep up with.

In addition, the cost of sequencing a genome is decreasing at a faster rate than is the cost of computation, although the curve seems to have been flattening out in the last 3–4 years (see Fig. 10).

We are thus faced with a scalability challenge – that is, 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.

To address the scalability challenge, we have recently begun to investigate mechanisms for improving the fault tolerance (with respect to network- and processor failures) of large parallel scientific software tools that run concurrently on thousands of cores by example of RAxML-NG, our tool for phylogenetic inference.

Another novel line of research in this area is our new focus on making such large computational codes more energy-efficient. Again, we conduct research in this domain by example of RAxML-NG as it is the most widely used and most scalable Bioinformatics tool developed in our group.

Hence, it also generates the largest CO<sub>2</sub> footprint. Initial experiments have revealed that using fewer cores for the computations and reducing the clock frequency of the cores (as our computations are predominantly memory-bandwidth-bound – i.e., the cores waste cycles/energy by waiting to retrieve data from the main memory) can substantially decrease the

total amount of energy-to-solution required.

Overall, phylogenetic trees (evolutionary histories of species) and the application of evolutionary concepts in general are important in numerous domains of biological and medical research. Programs for tree reconstruction that have been 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 – *inter alia* – to disentangle the origin of bacterial strains in hospitals, to determine the correlation between the frequency of speciation events (species diversity) and past climatic changes, and to analyze microbial diversity in the human gut.

Finally, phylogenies play an important role in analyzing the dynamics and evolution of the current SARS-CoV-2 pandemic and in conducting local contact tracing. To that end, some of our activities also focused on contributing to the analysis of the vast number of SARS-CoV-2 virus genomes that have been sequenced, which now total approximately 225,000. We describe some of these activities in greater detail in the following sections.

### Phylogenetic analysis of SARS-CoV-2 data is difficult!

The genomic data on SARS-CoV-2 exhibit several properties that render their phylogenetic analysis difficult. In a project that emerged ad hoc in spring 2020 [Morel, 2020], we decided to investigate whether it is possible to reliably reconstruct large phylogenetic trees on SARS-CoV-2 genome data.

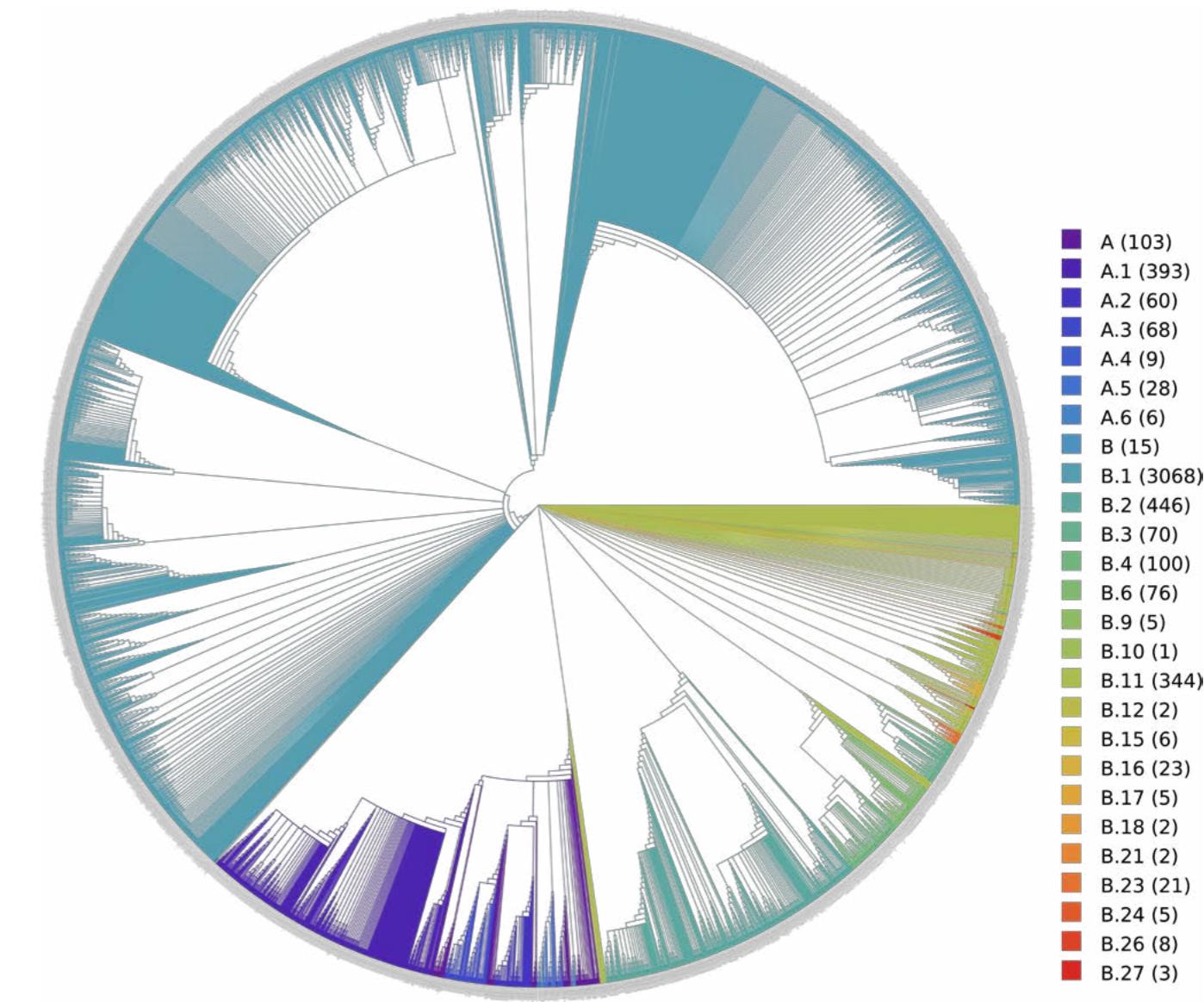
Together with our lab alumni Dora Serdari, Pavlos Pavlidis, and Lucas Czech as well as all current staff members of the lab and virus-evolution experts from Greece and Cyprus, we set forth to explore the difficulties of analyzing the evolution of these challenging data. Apart from the scientific outcome, frequent video conferences during the first lockdown also helped to prevent social isolation and were generally just fun. Using a snapshot of the available whole-genome data on 5 May, we investigated the difficulties of analyzing the data, which are due to uneven sampling/sequencing across countries, to a large variation in sequence-data quality across countries, and to the generally very low mutation rate of the virus, which renders the phylogenetic tree reconstruction challenging.

We found that the phylogenetic signal in the data is generally very weak (as expected) and that it is therefore not sufficient to reconstruct a single phylogenetic tree because the likelihood surface is extremely rugged. In other words, a large number of phylogenetic trees (evolutionary hypotheses about the evolutionary history of the virus) can be found that exhibit similar likelihood scores – that is, trees, that explain the data equally well and that we cannot distinguish using standard statistical significance tests for phylogenetics. The key problem is that these equally plausible trees exhibit substantial topological differences – that is, despite having similar likelihood scores, the actual tree topologies can be vastly different from one another. To alleviate this problem, we introduced the concept of a ‘plausible tree set’ that contains all equally likely trees and proposed

that the entire set be summarized and used for any downstream analyses and interpretations of the results. Despite the weak signal, we found that our approach of summarizing the information in the plausible tree set does yield helpful information. For

structure of the consensus phylogeny and the virus classification (see Fig. 11). Importantly, this study also found that current tools for likelihood-based phylogenetic inference have substantial numerical difficulties when applied

pandemic – cannot be determined with confidence using either the most-closely known bat- or pangolin coronaviruses or the first human SARS-CoV-2 genomes from the Wuhan region.



**Figure 11:** Consensus tree constructed from the plausible tree set of SARS-CoV-2 data available on 5 May, annotated by the virus-subtype classification. The different subtypes and their names are shown in the bar on the right.

instance, we mapped a current virus subtype classification onto the consensus tree constructed by summarizing the tree topologies in the plausible tree set, and we observed a substantial concordance between the

to such data, for instance, when estimating the branch lengths of the trees or when using more complex statistical models of evolution. In addition, we found that the root of the tree – that is, the starting point of the

## The Serratus Project

Our PhD student Pierre Barbera also became involved in a second project associated with a freely available open-source cloud-based computational analysis tool called Serratus that was designed primarily for studying coronavirus data [Edgar, 2020]. Pierre integrated the EPA-NG tool for phylogenetic placement of anonymous sequence data that he developed while working on his thesis as well as methods for standard phylogenetic inference into Serratus.

The overall goal of Serratus is to provide the infrastructure and methods required to carry out sequence-similarity-based searches with the goal of identifying closely related sequences to a set of query sequences, such as all genes from all available coronavirus sequences (this collection of all genes from all subtypes/strains of an organism/virus is also commonly denoted as the pangenome) in huge collections of public sequencing data. In this context, EPA-NG was deployed to taxonomically classify sequences that might belong to the broad spectrum of coronaviruses.

More specifically, Serratus was used to search and align the coronavirus pangenome against publicly available sequencing data containing 5.6 petabases of data (i.e.,  $5.6 \times 10^{15}$  nucleotides).

The computational analysis using Serratus helped to identify thousands of coronavirus- and coronavirus-like genomes and genome fragments in the reference data, including known subtypes as well as – more importantly – putatively novel types. Finally, the Serratus approach is highly generic – that is, it can also be used to unravel putative novel types in other

viruses, such as the delta viruses or the bacteriophages.

## CellPhy - a tool for inferring single-cell phylogenies

Another project related to the application of phylogenetic methods to public health problems is the development of the CellPhy tool [Kozlov, 2020]. Pierre integrated the EPA-NG tool for phylogenetic placement of anonymous sequence data that he developed while working on his thesis as well as methods for standard phylogenetic inference into Serratus.

Single-cell sequencing thus represents a revolution in modern biology.

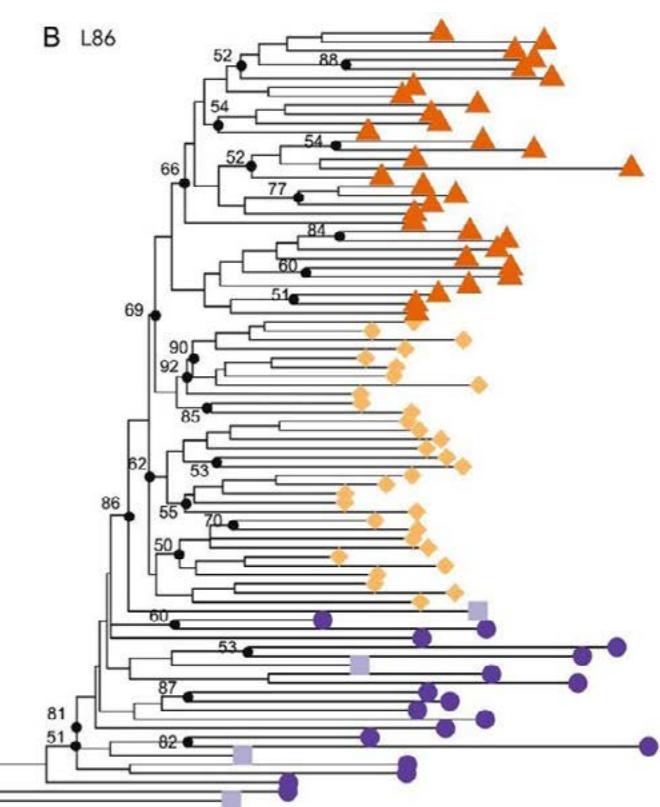
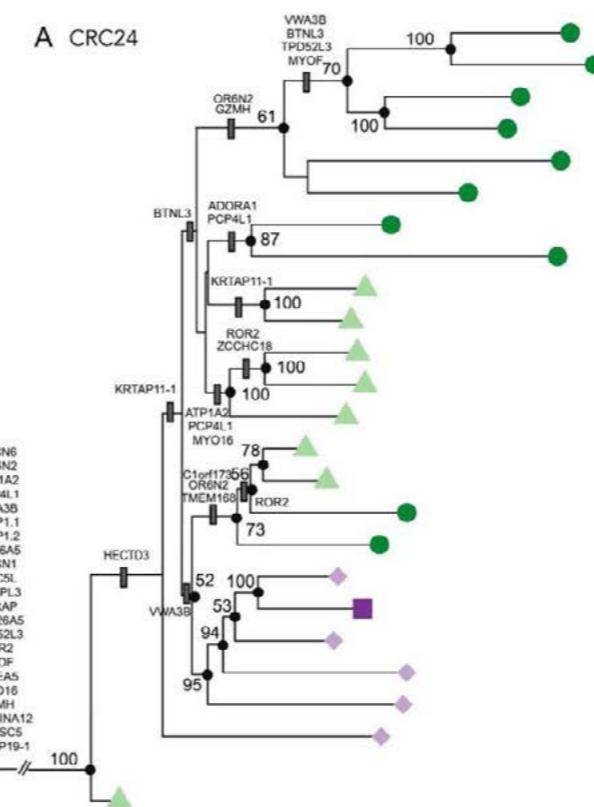
Beyond cancer, understanding this genomic diversity within a single living organism has applications in several areas of biology due to its connection with development and aging processes as well as with genetic diseases. The inference of phylogenetic trees on this novel type of data goes hand in hand with several challenges. For example, we needed to substantially adapt our statistical models of nucleotide evolution in order to account for the idiosyncrasies of single-cell genome data. In general, these data tend to be noisier than ‘normal’ DNA data used for standard phylogenetic inference and also contain more – as well as several distinct types of – sequencing errors.

To that end, we developed dedicated statistical models for the evolution of single-cell sequencing data, including an error model as well as a method to account for uncertainty in the input data. This novel model was implemented in the standard RAxML-NG tree search algorithm and tested on simulated single-cell sequence data under a large number of scenarios

(e.g., including/excluding errors, distinct rates of evolution, etc.).

Based on a total of 19,400 simulated datasets, our results reveal that CellPhy is robust with respect to the different error types and sources of uncertainty and that it outperforms competing methods under realistic simulation scenarios with respect to both tree accuracy and speed. We also applied CellPhy to a new empirical single-cell genome dataset for colorectal cancer and to three previously published empirical datasets. Figure 12 displays two trees that have been reconstructed on the genomes of healthy cells and tumor cells from different parts of the body of two colorectal cancer patients (A and B).

In the context of the emerging discipline of single-cell sequencing-data analysis, Alexis and Alexey also contributed to a review paper that summarizes the 11 main challenges faced by single-cell data science [Lähnemann, 2020].



**Figure 12:** Sub-figure (A) CellPhy tree for 24 colorectal cancer-cell genome sequences. The distinct shapes and colors at the tips of the tree denote the respective cell type. A healthy cell is depicted by a blue circle, whereas the cancer cells are depicted by other colors and shapes depending on the type of cancer cell and its location in the body of the patient. Sub-figure (B) CellPhy tree for 86 colorectal cancer-cell genome sequences. Healthy cells are denoted by dark purple circles and light purple squares. Tumor cells of different types are denoted by light orange diamonds and dark orange triangles.

Die Gruppe **rechnerbasierte Molekulare Evolution (CME)** beschäftigt sich mit Algorithmen, Modellen 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
- Diskrete Algorithmen auf Bäumen
- Methoden der Populationsgenetik.

In diesem Bericht 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“.

# 2 Research

## 2.4 Computational Statistics (CST)



### Group Leader

Prof. Dr. Tilmann Gneiting

### Staff members

Dr. Jonas Brehmer (since June 2020)

Dr. Timo Dimitriadis (until September 2020)

Dr. Alexander I. Jordan (staff scientist since July 2020)

Johannes Resin

Patrick Schmidt (until March 2020)

### Visiting scientists

Dr. Johannes Bracher (since May 2020)

Dr. Timo Dimitriadis (since October 2020)

Dr. Sebastian Lerch

Eva-Maria Walz

### Student

Daniel Wolffram (since September 2020)

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's research is focused on the theory and practice of forecasting. As the future is uncertain, forecasts should be probabilistic in nature, which means that they should take the form of probability distributions over future quantities or events. Accordingly, over the past several decades, we have been witnessing a trans-disciplinary shift of paradigms from deterministic- or point forecasts to probabilistic forecasts. The CST group seeks to provide guidance and leadership in this transition by

developing both the theoretical foundations for the science of forecasting and cutting-edge methodology in statistics and machine learning, notably in connection with applications. While weather forecasting and collaborative research with meteorologists continue to represent prime examples of our work, we have also addressed challenges raised by the pandemic in 2020 by establishing collaborative relationships with epidemiologists, creating the German–Polish COVID-19 Forecast Hub, developing methods for epidemiological ensemble forecasts, and contributing to similar efforts worldwide while placing methodological emphasis on the generation and evaluation of epidemiological ensemble forecasts.

### General news

There is no doubt that 2020 was a most unusual year for all of us. The year was shaped by a pandemic that has changed our lives in unprecedented ways. The CST group reacted quickly to the pandemic and contributed to the global effort on COVID-19 forecasting in various ways, including but not limited to the development of the German–Polish COVID-19 Forecast Hub. In parallel, we maintained an intense and fruitful interdisciplinary collaboration and exchange with meteorologists at our home university, the Karlsruhe Institute of Technology (KIT), and at the European Centre for Medium-Range Weather Forecasts (ECMWF) in Reading in the United Kingdom. Facets of this research are detailed in respective sections below. For the early-career researchers in our group, 2020 was a particularly rewarding year marked by much success and many accomplishments.

Johannes Bracher completed his PhD work in Epidemiology and Biostatistics at the University of Zürich in Switzerland in spring 2020 and began a postdoc position in Melanie Schienle's group at KIT, paired with an appointment at HITS. Immediately following his appointment, Johannes was included in KIT's competitive Young Investigator Group (YIG) Preparation Program and received start-up funding toward a junior research group.

Johannes also took the lead in developing the German–Polish COVID-19 Forecast Hub and has contributed to related efforts worldwide.

Jonas Brehmer completed his PhD work in Mathematics at the University of Mannheim and successfully defended his thesis in early December 2020, graduating with top honors. His PhD thesis, "Theory and Methodology of Scoring Functions: Tail Properties, Interval Forecasts, and Point Process-

es," addresses critical needs in evaluating earthquake forecasts and also deals with epidemiological forecasts in the interval format at the COVID-19 Forecast Hub.

Timo Dimitriadis was offered a prestigious new position as Assistant Professor in Applied Econometrics at Heidelberg University's Alfred Weber Institute of Economics, where he is currently based. Timo continues to be affiliated with HITS as a visiting scientist. Incidentally, in this new position, Timo is the successor to another HITS and CST alumnus, Fabian Krüger.

Our visiting scientist Sebastian Lerch received a highly competitive major four-year award in the "MINT for the Environment" program of the Vector Foundation to pursue research on "Artificial Intelligence for Probabilistic Weather Forecasting" at the cutting edge of the interface between mathematics, computer science, and meteorology. Sebastian is currently building a junior research group at our home university of KIT and maintains close ties to HITS.

Patrick Schmidt moved to a postdoc position at the University of Zürich in Switzerland, which began in April 2020. In July, we were pleased to welcome Alexander Jordan back into the group.

After completing PhD studies at HITS and KIT, Alexander moved to an academic position at the University of Bern in Switzerland. Back at HITS, in addition to expanding his research on statistics and machine learning, Alexander has been tackling important long-term tasks in the CST group, such as the continued development and maintenance of software tools that enable researchers across disciplines to apply our methodological solutions, particularly in the areas of distribution-

al regression and forecast evaluation. Alexander's support facilitates and invigorates our research in major ways and has already borne substantial fruit at this early stage, which will be detailed in the report for the coming year.

In continuation of a valued tradition, an integral aspect of our work is characterized by intense and frequent disciplinary and interdisciplinary scientific exchange. Before this exchange had to be shifted almost exclusively to an online format beginning in March 2020, we had organized an interdisciplinary event on the HITS premises in a joint endeavor with the Molecular Biomechanics (MBM) and Physics of Stellar Objects (PSO) groups. As reported on in Section 5.1.1, "Emulator Day" took place on 27 January 2020. Emulators are simple yet smart and computationally efficient surrogate models that complement more-complex, potentially prohibitively computationally expensive physics-based computer simulations. In the course of 2020, we strengthened the partnership between the three groups, and PhD students Kiril Maltsev and Kai Riedmiller joined HITS to work on mutual projects. While Kai and Kiril are based in the MBM and PSO groups, respectively, we were pleased to additionally welcome them to the CST group.

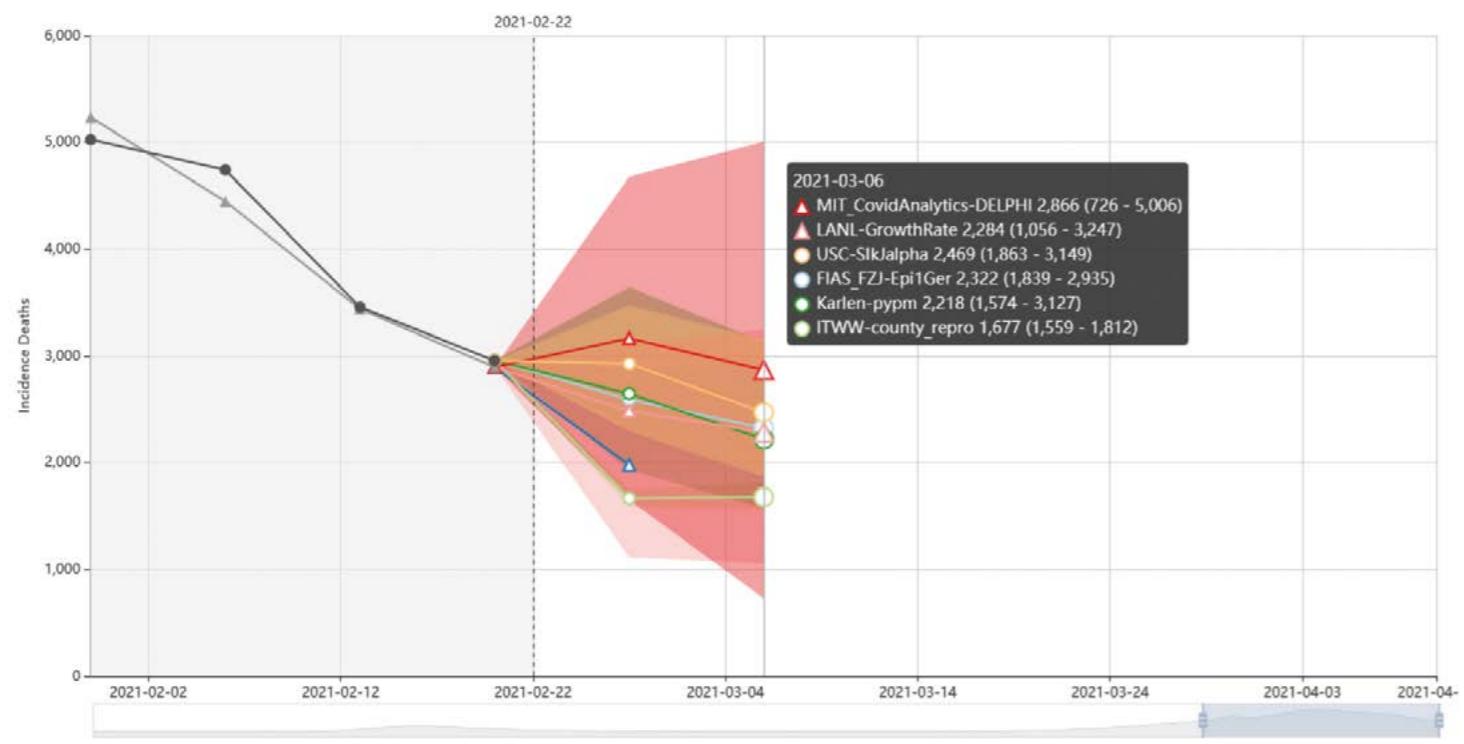


Figure 13: Screenshot from the interactive web application (<https://kitmetricslab.github.io/forecasthub/>) showing median forecasts and 95% forecast intervals for deaths in Germany from selected models at one and two weeks in advance, as issued on 22 February 2021.

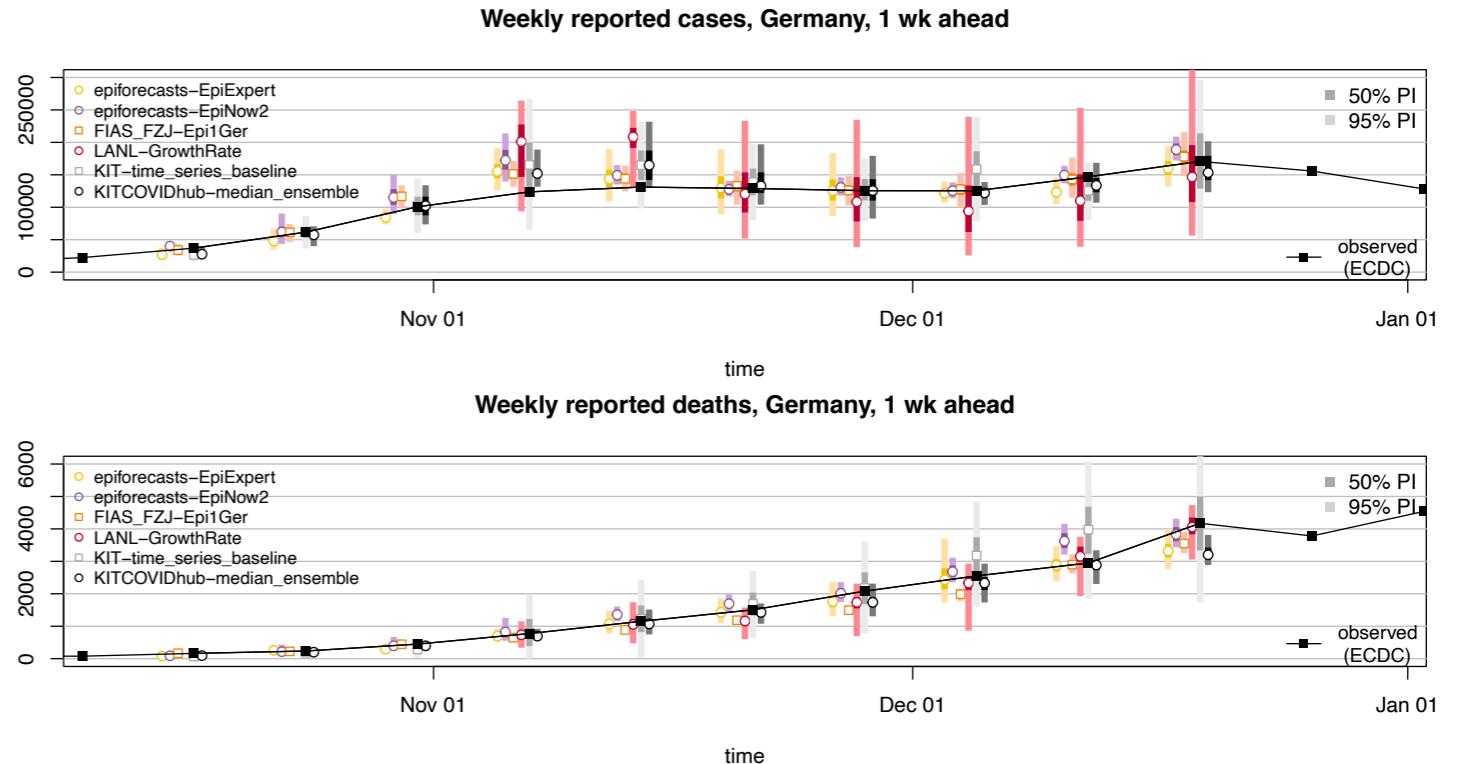


Figure 15: Forecasts one week in advance for cases (top) of and deaths (bottom) from COVID-19 in Germany from six selected models. We show 50% and 95% prediction intervals overlaid with data from the ECDC.

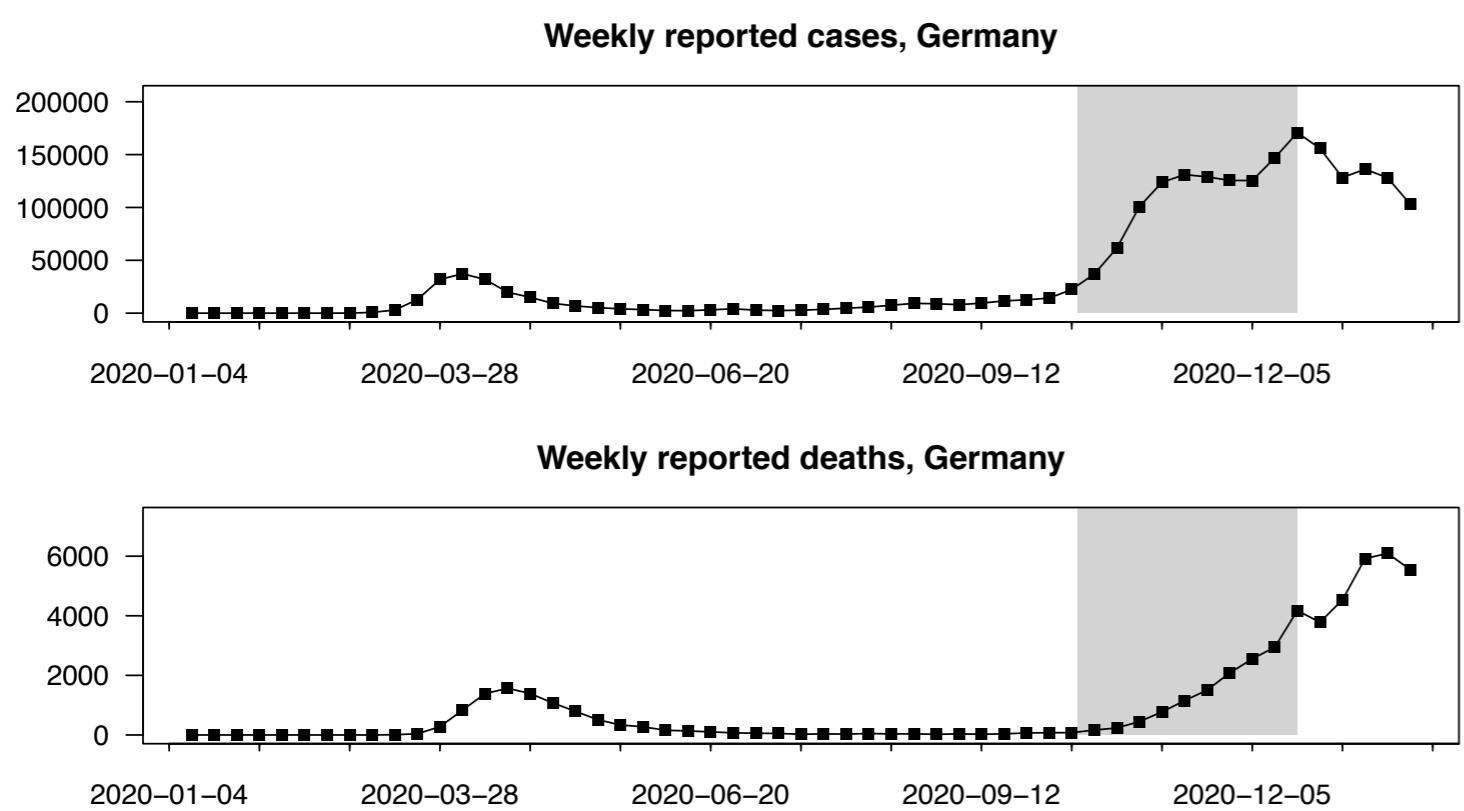


Figure 14: Weekly reported new cases (top) of and deaths (bottom) from COVID-19 in Germany according to data from the European Centre for Disease Prevention and Control (ECDC). The study period – from early October to mid-December 2020 – is highlighted in gray.

In what follows, we describe the development of the German–Polish COVID-19 Forecast Hub at KIT and HITS. Afterward, we move to collaborative research within the “Waves to Weather” consortium and report on a recent comprehensive study on the quality of precipitation forecasts for the tropics.

### German–Polish COVID-19 Forecast Hub

A highlight of our work in 2020 was the establishment of the German and Polish COVID-19 Forecast Hub (<https://kitmetricslab.github.io/forecasthub/>) in a joint endeavor with Melanie Schienle’s group at the Chair of Econometrics and Statistics in the Department of Economics at KIT with support from the Signale Team at the Robert Koch Institute. The Forecast Hub is run in close exchange with the US COVID-19 Forecast Hub (<https://covid19forecast-hub.org/>), to which we also contribute.

Johannes Bracher is a member of the US Hub team, and group leader Tilmann Gneiting serves on its Ensemble Advisory Board. Both Johannes and Tilmann are coauthors on preprints that report on results and experiences from the US Hub [Ray et al., 2020, <https://www.medrxiv.org/content/10.1101/2020.08.19.20177493v1>; Cramer et al., 2021, <https://www.medrxiv.org/content/10.1101/2021.02.03.21250974v1>].

In the German and Polish Forecast Hub, we collaborate with more than a dozen modeling teams from Germany, Poland, Switzerland, the United Kingdom, and the United States, who provide weekly forecasts of confirmed cases of and deaths from COVID-19 at prediction horizons of up to four weeks in advance in real time using a standardized format that consists of predictive quantiles at a total of 23 distinct levels. In addition, we produce and provide a range of baseline forecasts ourselves, and we aggregate

the various individual forecasts into a simple ensemble forecast that treats its members equally. All these types of forecasts are publicly available in an online repository (<https://github.com/KITmetricslab/covid19-forecast-hub-de>) and can be explored interactively in a dashboard, which is intended for a general audience and displays median forecasts along with forecast intervals. Figure 13 illustrates an example for deaths in Germany.

On 8 October 2020, we deposited a forecast-study protocol at the registry of the Open Science Foundation (OSF) that defined a study period and procedures for a prospective forecast-evaluation study. As Figure 14 illustrates, the specified study period spanned 10 weeks – from early October to mid-December – and coincided with the rise of the second COVID-19 wave in Germany. On 2 November, a national semi-lockdown was imposed, followed by a full

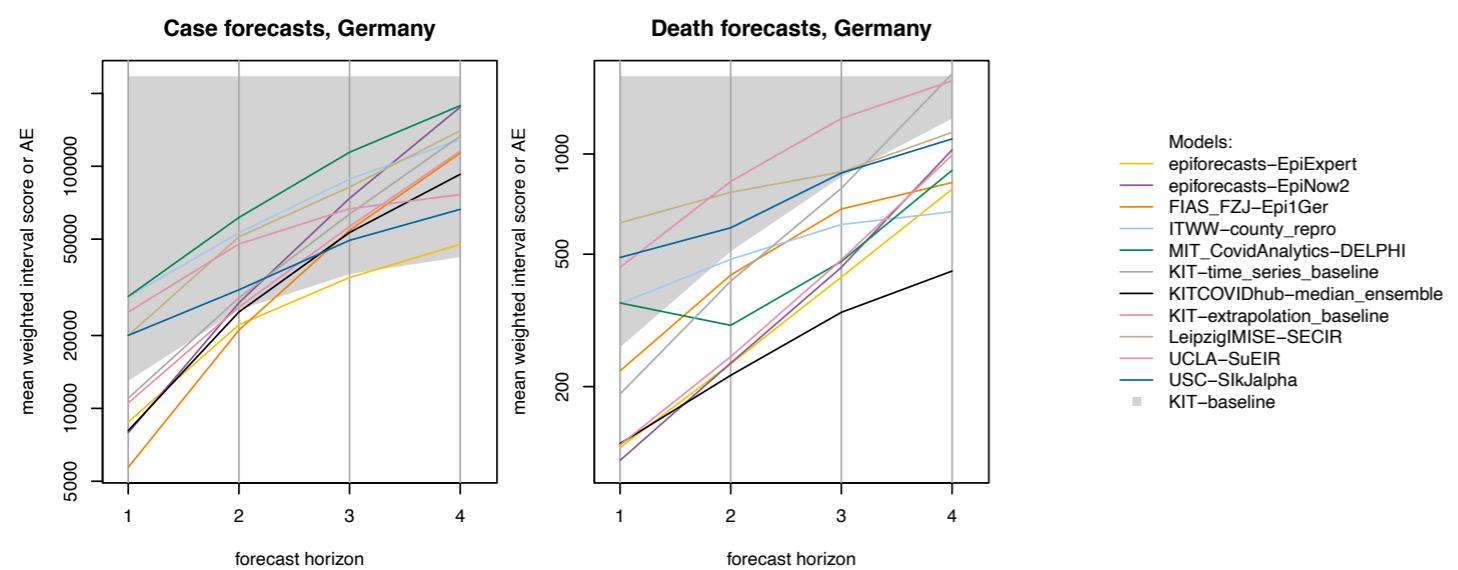


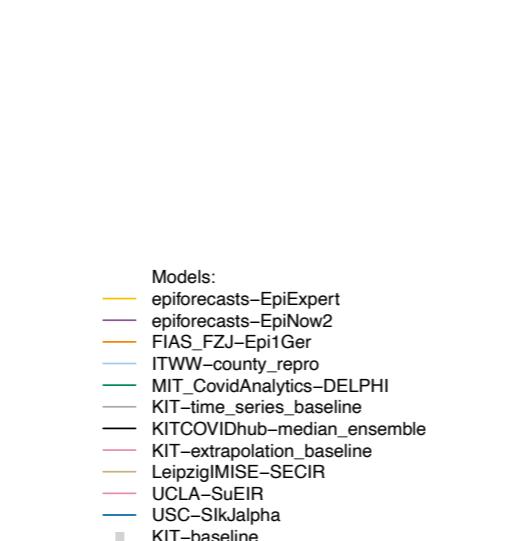
Figure 16: Mean weighted-interval score (WIS) achieved by various models for predictions of cases and deaths in Germany. Results are stratified by prediction horizon. For deterministic models, the mean absolute error (AE) is shown. The lower end of the gray area corresponds to the performance of the KIT-baseline model. Line segments within the white area thus indicate that a model outperforms the baseline.

lockdown from 16 December on. Figure 15 shows forecasts one week in advance for cases and deaths in Germany over the study period. While some forecast models account for interventions and changes in testing strategies, many groups prefer to use purely data-driven modeling strategies, which may entail delayed adaptation and explains why numerous models – including the ensemble forecast – showed overshoot in the first half of November, when cases began to plateau. For the final week of our study period, most models considerably underestimated deaths, perhaps due to prior age-specific shifts in incidence rates. Interestingly, the models differ from one another with respect not only to their point (median) forecasts but also to the implied uncertainty, as reflected by the 50% and 95% forecast intervals.

The forecasts were evaluated using the mean weighted-interval score (WIS), which is a theoretically coherent measure that reduces to the mean

absolute error (AE) in the case of a deterministic model that issues point forecasts only. Figure 16 provides an overview of results for Germany by target and forecast horizon. For cases, which are a more immediate measure of COVID-19 activity than deaths, hardly any approach succeeded in outperforming the naïve baseline forecast beyond a prediction horizon of two weeks. However, for deaths, both the ensemble forecast and a majority of the individual models outperformed the baseline at prediction horizons of up to four weeks in advance.

Overall, we found that achieving good probabilistic forecasts is challenging in a dynamic epidemic situation, and we observed pronounced heterogeneity between the different forecasts, with a general tendency toward overconfident forecasts, as evidenced by overly narrow prediction intervals. However, collaborative forecasting is beneficial, and our ensemble forecast displayed good – albeit not outstanding – relative performance.



## Toward better precipitation forecasts for the tropics

In Europe and North America, reliable weather forecasts have been taken for granted for decades. For many tropical countries, accurate forecasts of rainfall are unavailable despite the critical role this rainfall plays in the effective management of key resources, such as water, food, health, and renewable energy. In a recently published joint study with meteorologists at KIT [Vogel et al., 2020], we assessed the usefulness of predictions of 1- to 5-day accumulated precipitation from leading weather centers. Due to the sparsity of weather stations over large parts of the tropics, the study relies on pseudo-observations of rainfall from satellites.

In order to generate weather forecasts, weather centers draw on highly sophisticated numerical models that are run on supercomputers in real time and produce point forecasts of future atmospheric states. In a strong move toward probabilistic forecasts, these efforts have been transformed through the operational implementation of ensemble systems since the 1990s. An ensemble forecast consists of multiple simultaneous runs of numerical-weather-prediction (NWP) models, which differ from one another in terms of the two major sources of uncertainty, namely the initial state of the atmosphere and the mathematical representation of the respective physical processes. We focus on results for the world-leading 52-member ensemble operated by the ECMWF. Despite their

undisputed successes, ensemble systems continue to exhibit systematic errors, such as biases (the predictions are systematically too high or too low) and dispersion errors (the output is systematically too concentrated or too spread out), as illustrated in Figure 18 (see following page) for 1-day accumulated rainfall (see the following page). It is therefore common practice to statistically postprocess the NWP output, and we use a state-of-the-art ensemble-model-output-statistics (EMOS) technique for this postprocessing. The respective statistical parameters need to be estimated from training data based on forecast-observation pairs from the past.

As a benchmark, we use the concept of probabilistic climatology with rainfall

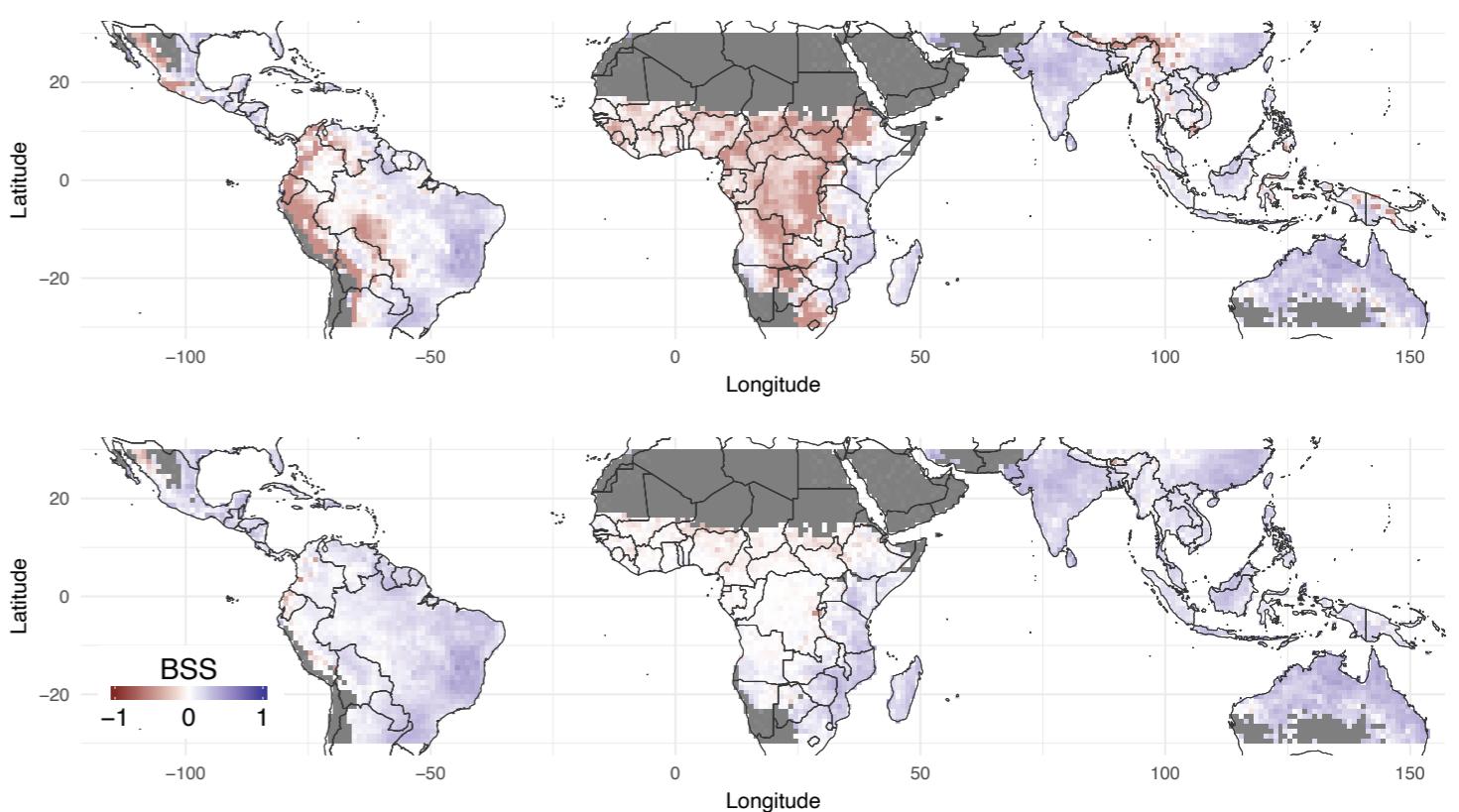


Figure 17: Brier skill score (BSS) for probability forecasts of 5-day rainfall accumulations of greater than 50 mm based on the raw (top) and statistically postprocessed (bottom) ECMWF ensemble relative to probabilistic climatology from 2009 to 2017. A skill score that is positive or negative corresponds to better or worse performance, respectively, than the reference forecast. Source: Vogel et al., 2020.

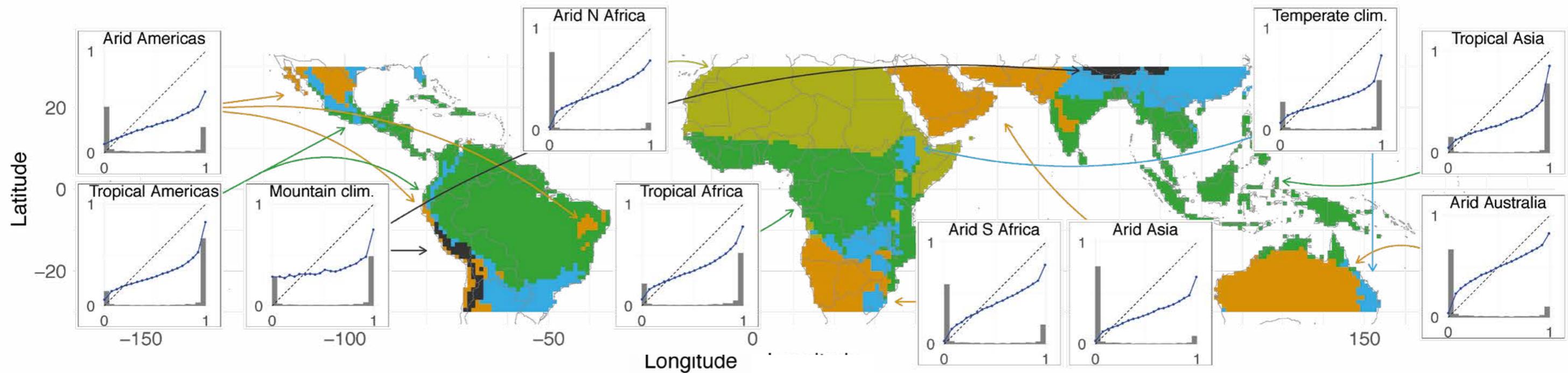


Figure 18: Calibration of 1-day ECMWF raw ensemble forecasts for precipitation accumulation over the tropics from 2009 to 2017. The probability-integral-transform (PIT) histograms correspond to the climate regions indicated with color shading and with the correspondingly colored arrows. For a calibrated forecast (i.e., a forecast that provides valid uncertainty quantification), the PIT histogram is uniform. The U-shaped histograms indicate that the raw ensemble is systematically too concentrated. Source: Vogel et al., 2020.

accumulations from a temporal window centered on the considered day of the year. This reference forecast represents the static, climatological distribution of rainfall at a given location and time of year but does not incorporate dynamic information about the state of the atmosphere. The raw and postprocessed ECMWF ensemble forecasts are clearly expected to outperform probabilistic climatology.

Putting this expectation to the test, Figure 17 (see previous page) shows the predictive performance of probabilistic forecasts for 5-day rainfall accumulations of greater than 50 mm in terms of the Brier score. A negative or positive skill score corresponds to a forecast that is inferior or superior, respectively, to the reference. Surprisingly, the raw ensemble forecast underperforms probabilistic climatolo-

gy in many parts of the tropics – a striking finding that remains true independently of accumulation time and weather center. Statistical postprocessing entails significant improvement of the raw model output in many regions, but not for most of tropical Africa, where performance remains sobering.

The fact that precipitation forecasts are poor for large parts of the tropics is a strong demonstration of both the complexity of the underlying forecast problem and the need for vigorous scientific development. Future work should refine statistical postprocessing methods, run NWP models at higher spatial resolution and with improved representations of physical processes, and merge NWP-based- and data-driven statistical forecasting techniques.

Die Forschungsgruppe **Computational Statistics (CST)** 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. Der Schwerpunkt der Forschung der Gruppe liegt in der Theorie und Praxis der Vorhersage.

Im Angesicht unvermeidbarer Unsicherheiten sollten Vorhersagen die Form von Wahrscheinlichkeitsverteilungen über zukünftige Ereignisse und Größen annehmen. Dementsprechend erleben wir seit nunmehr einigen Jahrzehnten einen trans-disziplinären 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 Methoden der Statistik und des maschinellen Lernens einnimmt und diese in wichtigen Anwendungsproblemen, wie etwa in der Wettervorhersage, zum Einsatz bringt.

In diesem Zusammenhang pflegen wir intensive Kontakte und Kooperationen mit Meteorolog/-innen am KIT und am Europäischen Zentrum für mittelfristige Wettervorhersagen (EZMW). Über kollaborative Projekte mit Epidemiolog/-innen, den Aufbau des deutsch-polnischen COVID-19 Forecast Hubs und die Unterstützung von ähnlichen Projekten weltweit stellen wir uns durch die Pandemie ausgelösten neuen Herausforderungen. Unsere besondere Aufmerksamkeit gilt dabei der Erzeugung und Bewertung von epidemiologischen Ensemblevorhersagen.

## 2 Research

### 2.5 Data Mining and Uncertainty Quantification (DMQ)



#### Group Leader

Prof. Dr. Vincent Heuveline

#### Staff members

Dr. Philipp Gerstner  
Alejandra Jayme (until August 2020)  
Philipp Lösel  
Dr. Chen Song

#### Visiting scientists

Saskia Haupt  
Sotirios Nikas (until June 2020)  
Elaine Zaunseder (since October 2020)  
Alejandra Jayme (since September 2020)  
Aksel Alpay (since November 2020)

#### Students

Charlotte Boys (until September 2020)  
Jonas Roller

The Data Mining and Uncertainty Quantification (DMQ) group, headed by Prof. Dr. Vincent Heuveline, began its research in May 2013. The group works in close collaboration with the Engineering Mathematics and Computing Lab (EMCL) at the Interdisciplinary Center for Scientific Computing (IWR) at Heidelberg University, which is also headed by Vincent Heuveline. The DMQ group's research focus lies in gaining knowledge from extremely large and complex datasets through data-mining technologies. Reliability considerations with respect to these datasets are addressed via methods of uncertainty quantification. Both fields – data mining

and uncertainty quantification – require a decidedly interdisciplinary approach to mathematical modeling, numerical simulation, hardware-aware computing, high-performance computing, and scientific visualization.

In 2020, the DMQ group focused on research activities in the areas of uncertainty quantification and machine learning for medical applications, mathematical oncology, and the Biomedisa online platform for biomedical image segmentation.

#### Machine learning for predicting employee absences

In order to remain competitive and effective in the market, companies seek to reduce costs and maximize profit. Employee absenteeism, whether justified or not, is a crucial factor in reaching this goal. Within the KIPROSPER project, we study this issue via machine-learning techniques to gain insights into employee absences. Specifically, we aim to predict absences given employee data, such as demographics, medical and behavioral history, self-reported job satisfaction, and work-structure information. Furthermore, we analyze which of these employee data are significant factors in this prediction.

The task is treated as a classification problem. Classification aims to approximate a mapping function from input-independent variables (features) to output-dependent variables (targets). In this study, the features are medical measurements, demographic data, work-structure data, self-reported life- and work characteristics, and health- or behavioral history, and the target is a label corresponding to a range of the number of absent days, which represents an absence-risk level.

In this work, three machine-learning models are studied: random forests (RF), support vector machines (SVM), and artificial neural networks (ANN). Random forests construct a multitude of individually trained decision trees. Their output is the label with most votes from the trees. Overfitting is minimized using different versions of the input data (bootstrap aggregation) and random subsets of features for each tree during training. Support vector machines aim to construct a hyperplane or a set of hyperplanes that separate the data into groups

with as much distance between them as possible. Artificial neural networks are composed of layers of connected nodes through which information travels from the input layer (the first layer) to the output layer (the last layer) via one or more (hidden) layers. The connections between nodes have weights that adjust as learning proceeds through several traversals of the layers. Artificial neural networks perform tasks without programmed rules or prior knowledge and instead automatically learn from the examples or the training data that they are given to process.

Prior to training the models, the training data are normalized to a common scale and then resampled to synthesize additional samples in order to balance the number of per-class labels. The metric used for performance measurement is accuracy (i.e., the ratio of correct predictions to total predictions). Cross-validation is employed to generalize the model evaluation. Experiments are

features influence the target in artificial neural networks, partial derivatives that represent variations of the output with respect to small changes in the inputs are obtained. These derivatives are then ranked in order to determine the degree to which the model is sensitive to each feature.

Results reveal that artificial neural networks have the best accuracy (77%), followed by random forests (72%) and support vector machines (62%). Artificial neural networks also take at least 6 times longer to train and perform better in almost all test cases, except when the training-data size is smaller, in which case random forests have slightly better accuracy (1–2%). Random forests also yield consistent results, with accuracy ranging from 68–74%. Support vector machines always perform worst among the models. Results also show that the set of training features significantly affects models' performance. Highly correlated features

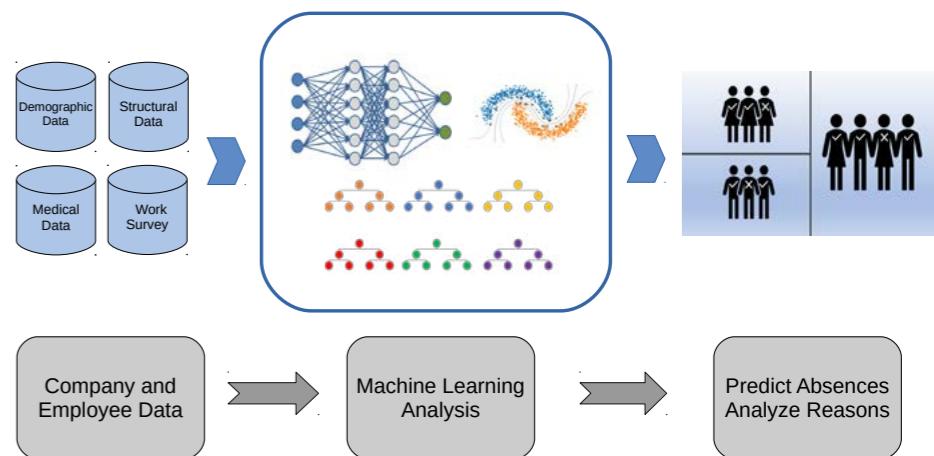


Figure 19: Machine learning for predicting employee absences.

designed to analyze the effect of training-data size, model size, and the number of features on the models' performance. The feature sets are varied by degree of correlation or by ranking according to a domain expert, feature importance, and random ranking. To analyze how input

contain the same information, and removing redundant features has little impact on the models' accuracy. Neural networks can also correctly classify the change in the risk of absence after switching the values of the features to which the model is highly sensitive.

These findings make clear that machine-learning methods can be used to successfully predict employee absences and similar health-related problems. A thorough data preparation with appropriate model selection is however the needed ingredient for reliable predictions.

### Uncertainty quantification in a multi-physics heart simulation

The contraction of a human heart can be described by a system of partial differential equations that combine hydrodynamical, elastomechanical, and electrophysiological elements. Understanding the underlying physical processes allows various cardiovascular diseases to be diagnosed and treated. In a joint project with the respective groups led by Christian Wieners, Bettina Frohnau, and Olaf Dössel, all from the Karlsruhe Institute of Technology, we address the issue of quantifying uncertainties that arise in the context of a multi-physics heart simulation.

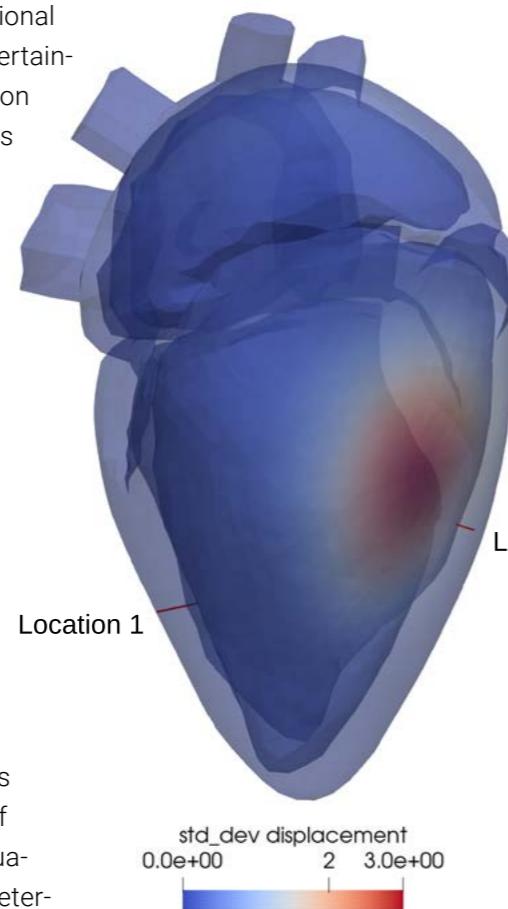
The studied physical model is based on the nonlinear equations of continuum mechanics and describes the displacement of the initial heart geometry under the influence of a force that arises due to biochemical reactions in the cells of the heart. As this situation is highly patient-specific and it is often not possible to measure material parameters in living organs with accurate precision, it is important to consider that the input data of the simulation are subject to uncertainties.

Examples of these uncertain parameters include material parameters, the blood pressure inside the heart ventricles, the alignment of muscle fibers, and the overall geometry that is obtained by imaging methods,

such as MRI. In order to quantify how these uncertainties affect the validity of given quantities of interest, we developed a computational framework that propagates the input uncertainties through a deterministic model. From a mathematical point of view, this framework is based on the generalized polynomial chaos method (gPC) and allows the arising random variables to be expanded in a series of orthogonal polynomials. The gPC discretization technique is combined with two additional algorithmic concepts in order to deal with the so-called curse of dimension, which states that the computational effort for uncertainty quantification (UQ) increases exponentially with the number of considered, uncertain parameters. The first concept in use is given by the Smolyak sparse-grid method, which reduces the number of required evaluations of the deterministic model – that is, the number of deterministic simulations for a given set of input parameters. The second concept is a multilevel technique. Here, the idea is to consider a hierarchy of deterministic models that is obtained by changing the underlying discretization parameters, which determine the accuracy and computational effort for performing a single deterministic

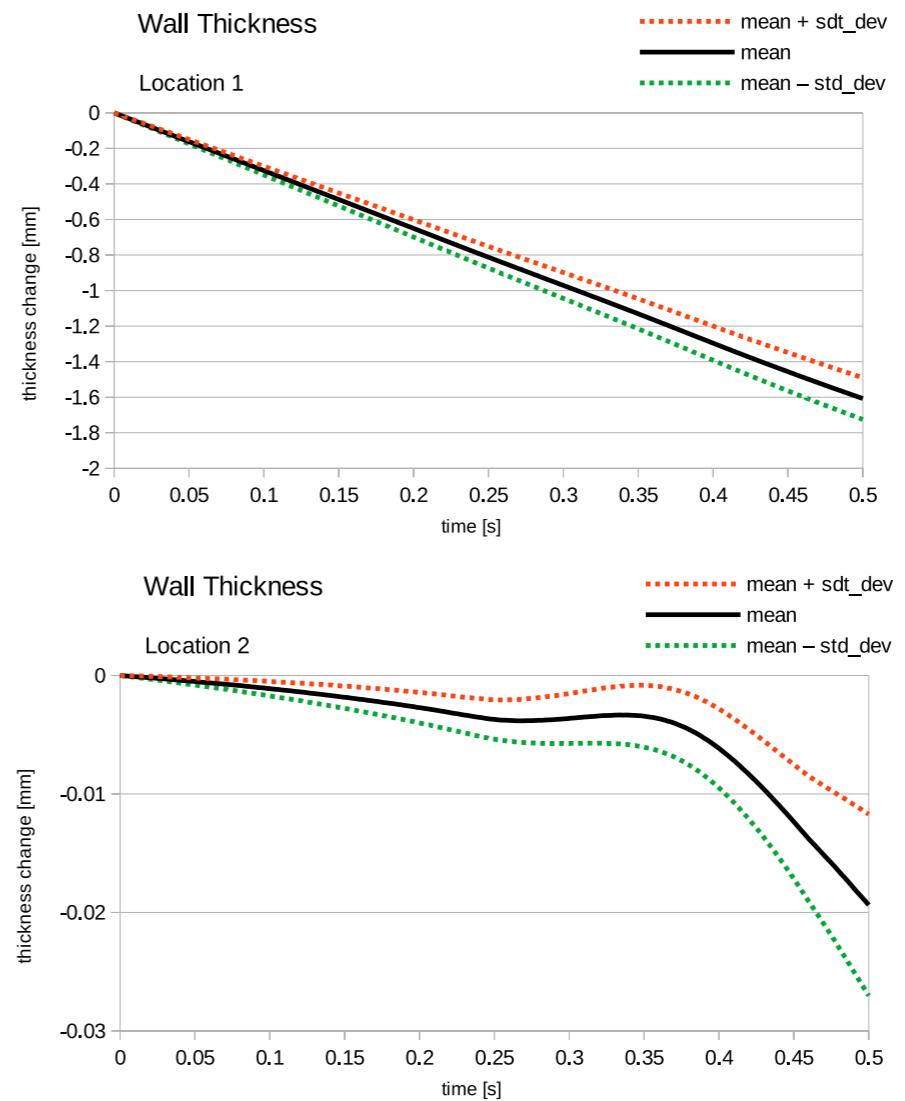
simulation. In a similar fashion, a hierarchy of Smolyak sparse grids is considered that is obtained by varying the number of grid points in the input-parameter space – that is, the number of deterministic evaluations. Next, we combine the gPC expansions that are obtained for high-accuracy sparse grids and low-accuracy deterministic simulations with those obtained for low-accuracy sparse grids and high-accuracy simulations in a defect-correction fashion. In this way, a favorable relationship between accuracy and computational effort is achieved.

Figure 20 illustrates the results of the previously described UQ workflow for uncertain material parameters and internal blood pressure. The figure shows the heart geometry at the time of maximal contraction during a single heartbeat. Here, the color map encodes the standard deviation of the solid



*Figure 20: Standard deviation of displacement field during heart contraction.*

displacement field. Regions of high uncertainty (i.e., regions in which the uncertain input parameters have a large effect on the displacement) are plotted in red (Location 2). In these regions, the validity of the physical



*Figure 21: Mean and standard deviation of wall thickness during heart contraction at two prescribed locations.*

model may be low, which must be taken into account in practical scenarios (e.g., by a surgeon who plans an operation based on simulation data). The plots in Figure 21 reveal how the wall thickness of the heart changes over time at the prescribed locations. At both locations, the muscle is stretched, as is indicated by negative values. At Location 2, however, the corresponding standard deviation is significantly larger (in relation to the mean value) than at Location 1. This behavior again illustrates the heterogeneous influence of uncertain input parameters. This topic is embedded in the Informatics4Life project (see Chapter 6).

### From semi-automatic to fully automatic image segmentation with the Biomedisa online platform

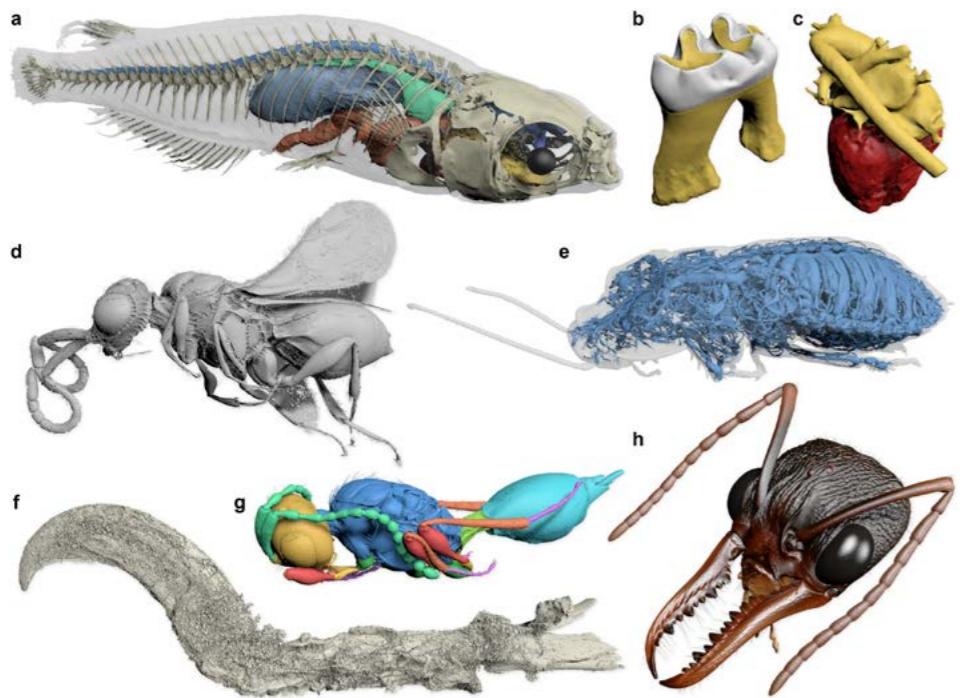
Three-dimensional imaging is leading to progress in many scientific disciplines. The analysis of volumetric medical and biological imaging data – for example, from X-ray computed tomography (CT), magnetic resonance imaging (MRI), or optical microscopy – often requires isolating individual structures from the 3D volume via segmentation. Ongoing improvements in imaging technologies result in higher resolutions and faster acquisition times, thereby

increasing the demand for accelerated image analysis. Image segmentation, in particular, remains a major bottleneck and is often the most labor-intensive and error-prone task of 3D-image analysis.

In situations in which little a priori knowledge is available, fully automatic segmentation routines are less feasible. In these cases, manual segmentation by an expert followed by morphological interpolation remains a very common approach. Here, labels are assigned to various structures of interest with different intervals inside the 3D volume (depending on the complexity of the dataset), followed by an interpolation of the labels between the pre-segmented slices. The underlying image data are usually not taken into account, and the interpolation is therefore based exclusively on the segmented slices. Consequently, only a fraction of the real experimental information is utilized to derive the segmentation.

To ensure the proper 3D segmentation of complex samples based on the morphological interpolation of pre-segmented 2D slices, dense pre-segmentation is required, sometimes even slice-by-slice. The conventional approach to the manual segmentation of 3D images is therefore often tedious and time-consuming, effectively impeding the analysis of large numbers of datasets from samples with high morphological variability, as required, for example, for digitizing scientific collections or for conducting quantitative studies on biodiversity. Furthermore, artifacts resulting from the morphological interpolation of manually segmented slices and subsequent correction (e.g., line artifacts, overly smooth meshes, etc.) limit the quality of the results.

## 2.5 Data Mining and Uncertainty Quantification (DMQ)

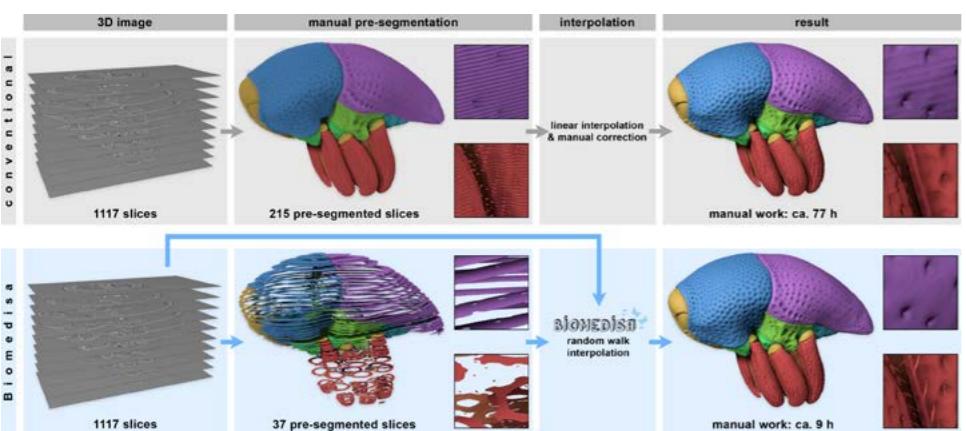


**Figure 22 : Biomedisa examples.** a) Medaka fish with a segmented skeleton and selected internal organs (based on  $\mu$ CT scan). b) Mouse molar tooth showing enamel (white) and dentine (yellow) ( $\mu$ CT). c) Human heart with segmented heart muscle and blood vessels (MRI). d) Fossilized parasitoid wasp from Baltic amber (SR- $\mu$ CT). e) Tracheal system of a hissing cockroach ( $\mu$ CT). f) Claw of a theropod dinosaur from Burmese amber (SR- $\mu$ CT). g) Fossilized parasitoid wasp preserved inside a mineralized fly pupa (SR- $\mu$ CT). h) Head of an Australian bull-ant queen (SR- $\mu$ CT). (Lösel et al., Nat. Commun. 2020).

With the goal of reducing the necessary effort required of a human annotator when segmenting large and complex samples of unknown composition, we developed Biomedisa (Biomedical Image Segmentation App; <https://biomedisa.org>), an easy-to-use open-source online platform that is accessible through a web browser and does not require any software installation or maintenance when used online. Biomedisa aims to be a one-button solution to addressing the needs of scientists without the need for complex and tedious configuration or for substantial computational expertise. Its semi-automatic segmentation is based on a smart interpolation of sparsely pre-segmented slices that takes into account the complete underlying image data. The method works well without parameter optimization and was specifically developed to help massively parallel computer architectures – such as graphics processing units (GPUs) – cope with constantly growing image sizes.

Biomedisa can be used for different 3D-imaging modalities and various biomedical applications (Figure 22). Its smart interpolation can vastly accelerate the segmentation process while simultaneously providing more-accurate results than manual segmentation (Figure 23).

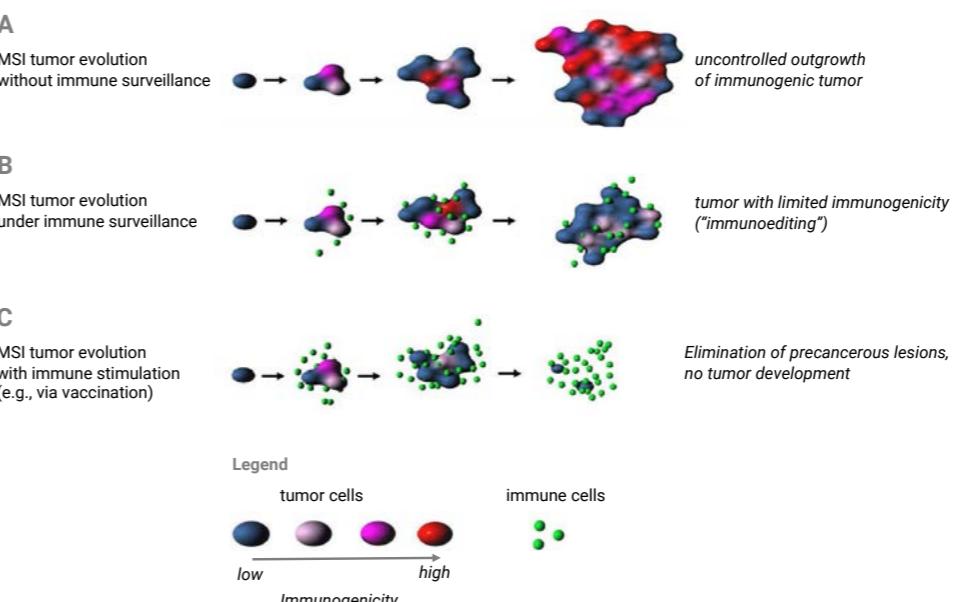
Finally, the results of the semi-automatic segmentation can be used to train a deep neural network that enables fully automatic segmentation when segmenting a large number of similar structures, such as the brain of insects or the human heart. As a result, Biomedisa enables large-scale comparative quantitative analyses to be used to gain a deeper understanding, for example, of the mechanisms behind insect cognition, and to provide numerical simulations based, for example, on a patient-specific heart model that supports clinicians in their surgical planning and decision-making.



**Figure 23: Comparison of a conventional segmentation approach (top row) and Biomedisa (bottom row).** The conventional procedure requires 77 hours compared with 9 hours with Biomedisa. Both procedures require manual pre-segmentation of the 3D-image stack. While the widely used morphological interpolation solely considers labels on pre-segmented slices, Biomedisa's random walk interpolation takes both the underlying 3D-image data and the pre-segmented slices into account, resulting in significantly less required manual input. Moreover, interpolation artifacts are avoided, and fine details – such as hairs, which are usually omitted during manual segmentation – are included (Lösel et al., Nat. Commun. 2020).

## The use of mathematical oncology to support the search for vaccines against cancer

Cancer is caused by various types of genomic instability. Among other instabilities, so-called DNA mismatch repair- (MMR)-deficient tumors account for about 15% of colorectal cancers and up to 30% of endometrial cancers. Notably, tumors that develop in the context of Lynch syndrome, the most common inherited cancer predisposition syndrome, are characterized by MMR deficiency. Approximately 1 in every 180 people



**Figure 24: Immune surveillance and immunoediting in mismatch repair-deficient, microsatellite-unstable cancers.**

may be affected by Lynch syndrome, which is associated with a high lifetime cancer risk. Knowing whether a tumor is MMR-deficient may help in planning treatment or predicting how well the tumor will respond to treatment.

MMR-deficient cells lack the ability to recognize and correct small errors in DNA during cell division, thereby leading to the accumulation of an exceptionally high load of mutations, particularly insertions or deletions (indels) of single nucleotides at repetitive sequences (microsatellite instability, MSI). Whenever gene-encoding microsatellites are hit, the respective gene function may be

abrogated, and frameshift peptides – novel protein structures caused by shifts of the translational reading frame – are generated. Such frameshift peptides can be recognized by the immune system, and MSI tumors may thus respond well to immune therapies. However, until now, it had not been known whether these neoantigens occur randomly in MSI cancers.

As part of the "Mathematics in Oncology" collaborative initiative with the Department of Applied Tumor Biology (ATB) at Heidelberg University Hospital, we developed and used a novel algorithm to quantitatively

detect microsatellite indel mutations with high sensitivity [Ballhausen et al., 2020]. In close collaboration, we identified mutations shared by most MSI colorectal and endometrial cancers. Such a predictable set of shared indel mutations is unique to MMR-deficient cancers and encourages the development of preventive vaccine approaches.

Furthermore, we detected a negative correlation between the prevalence of a defined indel mutation in MMR-deficient colorectal or endometrial cancers and the predicted immunogenicity of the resulting frameshift peptide. Our study strongly supports the concept of continuous immunoediting in human cancers (Figure 24B), which means that the immune system monitors the tumor during its development and immediately eliminates cancer cells with highly immunogenic neoantigens. The study provides new evidence for the hypothesis that immunogenic cancers and pre-cancer cell clones can be attacked and potentially eradicated by the host's immune system (Figure 24C). This finding strongly encourages the development of cancer-preventive vaccines that may help to reduce tumor risk in Lynch syndrome and potentially beyond.

Die Forschungsgruppe **Data Mining and Uncertainty Quantification (DMQ)** unter der Leitung von Prof. Dr. Vincent Heuveline besteht seit Mai 2013. Sie arbeitet eng mit dem „Engineering Mathematics and Computing Lab“ am Interdisziplinären Zentrum für Wissenschaftliches Rechnen der Universität Heidelberg zusammen, welches auch von Vincent Heuveline geleitet wird. Im Fokus der Forschungsarbeit steht ein zuverlässiger und strukturierter Wissensgewinn aus großen, komplexen Datensätzen, der mittels Data-Mining Technologien erreicht und mit Methoden der Uncertainty Quantification validiert wird. Beide Themenfelder – Data Mining und Uncertainty Quantification – erfordern Interdisziplinarität in den Bereichen mathematische Modellierung, numerische Simulation, hardwarenahe Programmierung, Hochleistungsrechnen und wissenschaftliche Visualisierung. 2020 wurde dazu in der Gruppe in folgenden Anwendungsbereichen gearbeitet: Uncertainty Quantification und maschinelles Lernen für medizinische Anwendungen, mathematische Onkologie und die Biomedisa-Webanwendung zur Segmentierung biomedizinischer Bilddaten.

# 2 Research

## 2.6 Groups and Geometry (GRG)



### Group Leader

Prof. Dr. Anna Wienhard

### Staff members

Johannes Horn (until October 2020)

Dr. Brice Loustau (since August 2020)

Mareike Pfeil

### Visiting scientists

Giulio Belletti (since October 2020)

Dr. Nguyen-Thi Dang

Dr. Valentina Disarlo

The Groups and Geometry research group works closely with the Research Station Geometry & Dynamics at Heidelberg University. Both groups are headed by Prof. Dr. Anna Wienhard.

Symmetries play a central role in mathematics as well as in other natural sciences. Mathematically, symmetries are transformations of an object that leave it unchanged. They can be composed – that is, applied one after the other – and form a mathematical structure called a group. In the 19th century, mathematician Felix Klein proposed a new definition of geometry as

### Marta Magnani

Dr. Andreas Ott (until September 2020)

Dr. Beatrice Pozzetti

Anja Randecker (since October 2020)

Evgenii Rogozinnikov

Dr. Carmen Rovi (since September 2020)

Dr. Andrew Sanders (until December 2020)

Anna Schilling

Dr. Gabriele Viaggi (since October 2020)

### Students

Levin Maier (since October 2020)

Menelaos Zikidis

the study of all properties of a space that are invariant under a group of transformations. In short: Geometry is symmetry.

This concept unified classical Euclidean geometry, the newly discovered hyperbolic geometry, and projective geometry, which has its origins in the study of perspective in art and is not based on the measurement of distances but rather on incidence relations. Klein's concept fundamentally changed our view of geometry in mathematics and theoretical physics and continues to influence these fields to this day.

In our research group, we investigate various mathematical problems in the fields of geometry, topology, and dynamics that involve the interplay between spaces – such as manifolds or metric spaces – and

groups, which act as symmetries of these spaces. We also apply the study of groups and geometry to other sciences, such as mathematical physics or data science and machine learning.

### Hyperbolic geometry

Hyperbolic geometry is the star of non-Euclidean geometries and provides the model for negative curvature. The long history leading to its discovery originates in Euclid's *Elements*, a monumental treatise of mathematics written in ca. 300 BC that is likely the most influential textbook ever written. Euclid's approach is axiomatic and constructive, relying on five fundamental axioms, which he calls *postulates*. For centuries, mathematicians questioned Euclid's fifth postulate, according to which, given a line and a point not on it, there exists a unique parallel line through the point (see Fig. 25 for Euclid's original formulation).

In the romantic 19th century, Lobachevsky, Gauss, Beltrami, Klein, Poincaré, and other mathematicians finally broke Euclid's rule: They developed a

consistent geometry in which the parallel postulate fails. This revolutionary discovery had profound consequences on mathematics, physics, and even philosophy. Hyperbolic geometry continued to play a prominent role through the 20th century, culminating in Thurston's geometrization program and its completion in the 2000s, which solved the renowned Poincaré conjecture, one of the seven Millennium Prize Problems. To this day, hyperbolic geometry and its derivatives remain an intensely active field of mathematical research. In this year's HITS report, we describe some of the main features and applications of hyperbolic geometry in connection with the research activities of the GRG group.

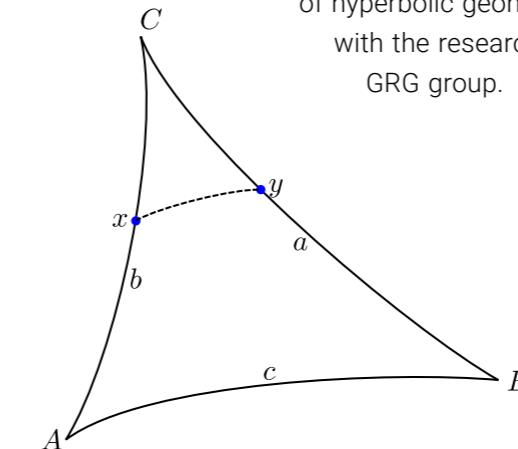


Figure 25: Euclid's original formulation of the parallel postulate: If  $\alpha + \beta$  is less than two right angles ( $180^\circ$ ), then the lines  $l$  and  $l'$  must intersect (Picture: Brice Loustau).

### Hyperbolicity in metric spaces

There are many equivalent approaches to describing hyperbolic geometry and the manifestation of negative curvature. One such approach is to compare the measurements of the lengths and angles in triangles – that is, trigonometry – in Euclidean versus hyperbolic geometry. For instance, hyperbolic triangles with given side lengths are always thinner than their Euclidean counterparts (see Fig. 26).

This type of observation led mathematicians to define curvature in very general metric spaces, which include

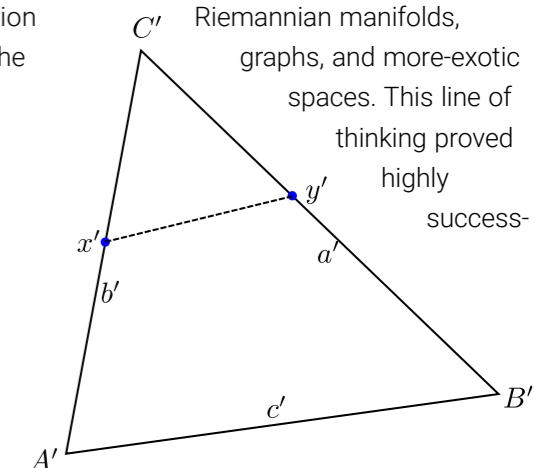
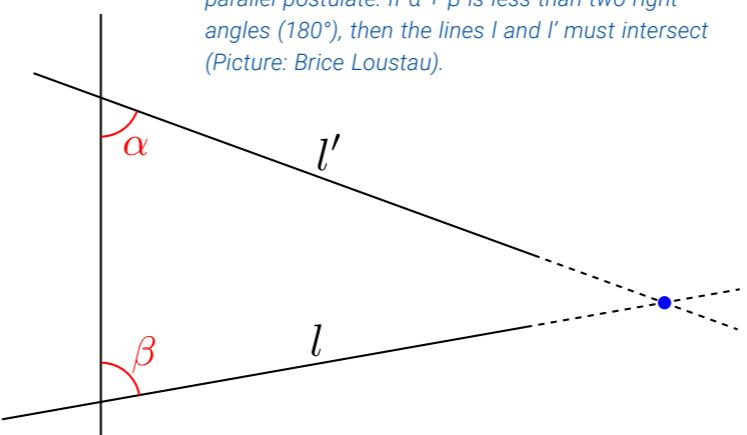


Figure 26: A hyperbolic triangle and its Euclidean counterpart. Hyperbolic triangles are "slim": Assuming the same side lengths  $|a| = |a'|$ ,  $|b| = |b'|$ ,  $|c| = |c'|$ , we have  $|x'y| < |x'y'|$ . (Picture: Brice Loustau).

Figure 25: Euclid's original formulation of the parallel postulate: If  $\alpha + \beta$  is less than two right angles ( $180^\circ$ ), then the lines  $l$  and  $l'$  must intersect (Picture: Brice Loustau).



ful in the study of discrete groups and gave rise to a sub-field of research known as *geometric group theory*, in which several GRG members specialize. One key idea of geometric group theory is to associate to any abstract group a geometric incarnation of it – called its *Cayley graph* – and to study this object with concepts and tools stemming from (hyperbolic) geometry, such as curvature and the classification of isometries according to their dynamics (see Fig. 27 on the following page).

## Representations of discrete groups and geometric structures

Another approach to hyperbolic geometry, in the spirit of Klein's "Geometry is Symmetry" (see above), consists of investigating the group of isometries of hyperbolic space. By definition, an isometry is a transformation of hyperbolic space that preserves its geometry (it is enough to require that it be distance-preserving). The discrete subgroups of isometries, such as groups of symmetries of tessellations of the hyperbolic plane, are particularly interesting (see Fig. 28).

More generally, the study of representations of discrete groups as isometries of non-Euclidean symmetric spaces is a highly active field of research that is closely related to the theory of geometric structures on manifolds and their deformation spaces. This field of research, which is sometimes called (higher) Teichmüller–Thurston theory, extends the classical Teichmüller theory of the deformation of complex structures on surfaces. Many GRG members specialize in this field (e.g., the 2020 publications [Pozzetti et al., 2020] and [Dang and Glorieux, 2020] fall into this category).

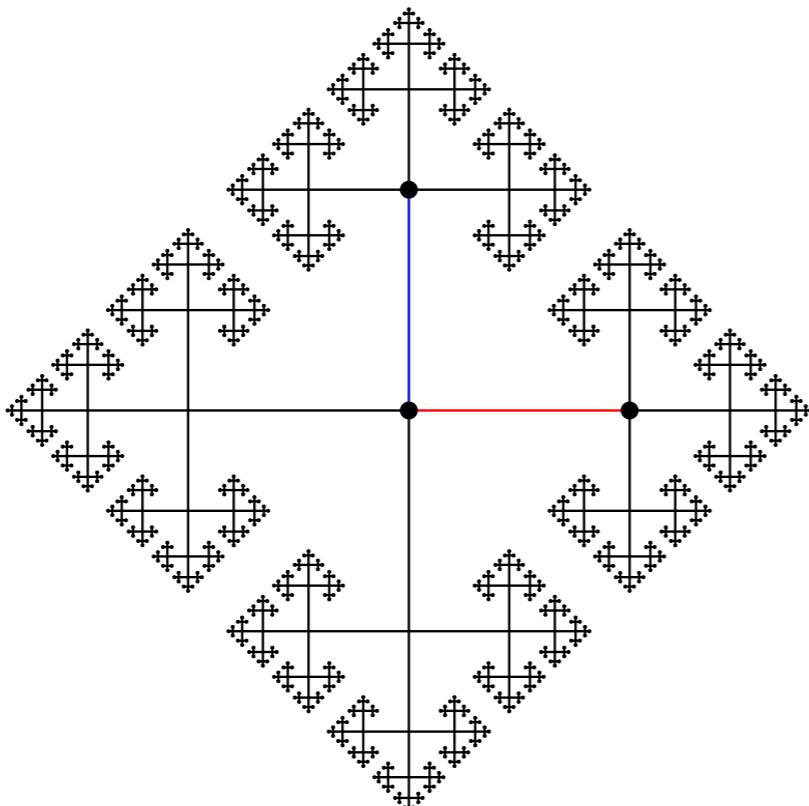


Figure 27: Cayley graphs of the free group on two generators and of a finite group of order 18. (Pictures: Brice Loustau, Wikipedia).

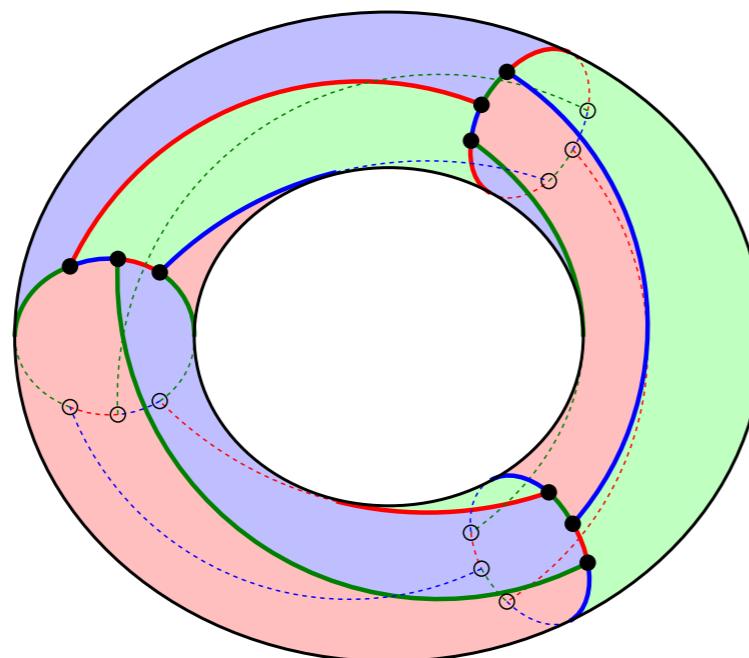
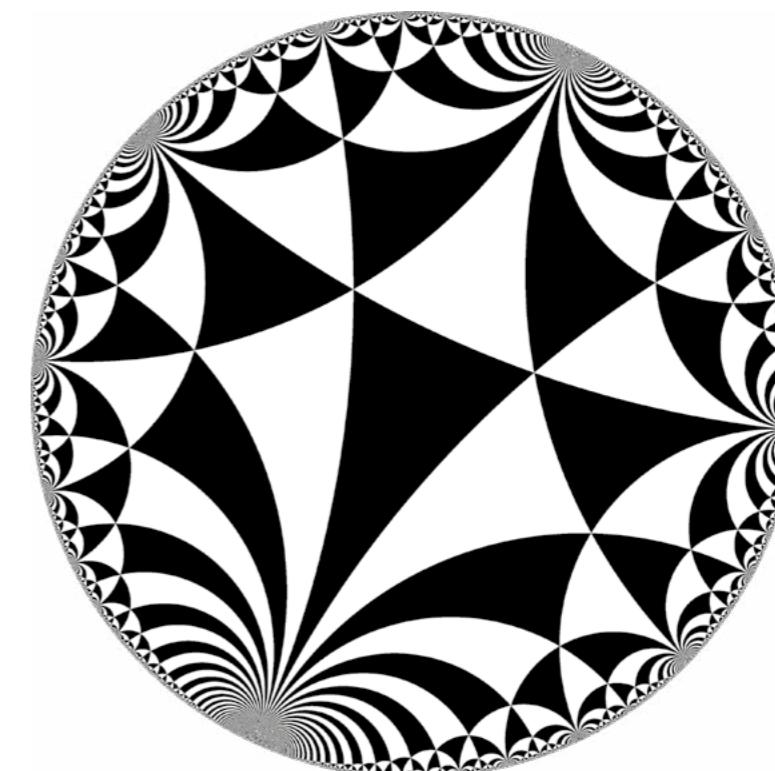
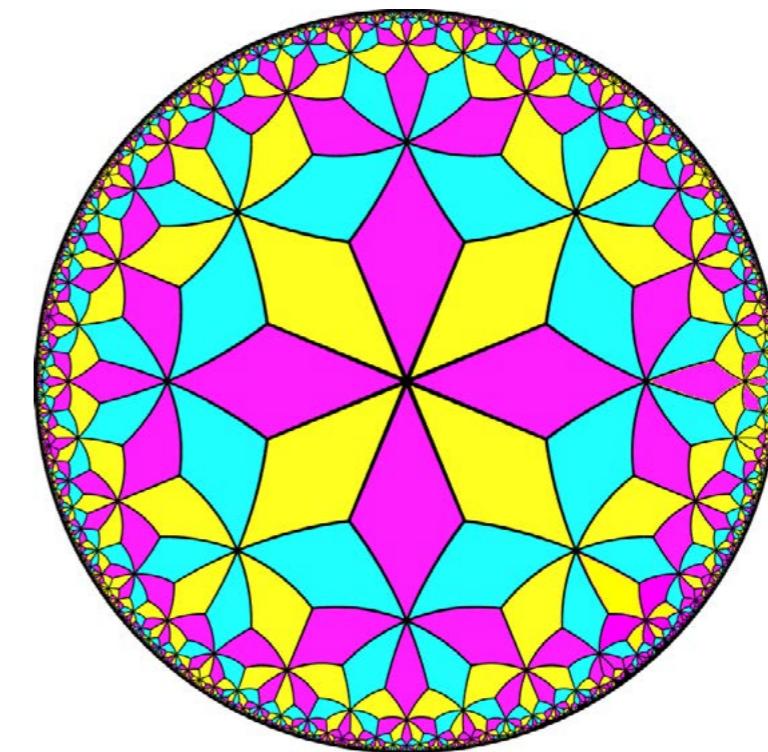


Figure 28: Two examples of uniform tilings of the hyperbolic plane in the Poincaré disk model, (Pictures: Wikipedia).



## Differential and computational geometry

The development of differential geometry based on the work of Gauss and Riemann tied the development of hyperbolic geometry to Riemannian geometry and to the study of certain differential equations, such as the geodesic equations and the Dirichlet problem. More generally, many problems from physics translate as PDEs (partial differential equations) on Riemannian manifolds, which mathematicians often attempt to solve using geometric methods. A useful tool in this field called *geometric analysis* consists of discretizing the domain and the equations and converting differential problems to discrete ones that are more amenable to computation and sometimes also to theoretical analysis. As an illustration, Fig. 29 features two examples of a harmonic map between two hyperbolic surfaces lifted to an equivariant map of the Poincaré disk with respect to the action of two Fuchsian groups. This map solves a nonlinear PDE that is the hyperbolic analog of the Dirichlet problem and was computed via a discrete heat-flow method (see [Gaster et al., 2019, Computing harmonic maps between Riemannian manifolds, arXiv preprint: 1910.08176] for details).

## Hyperbolic geometry and number theory

A deep connection exists between hyperbolic geometry and number theory that dates back to Poincaré's study of Fuchsian groups and automorphic forms relating to 2-dimensional hyperbolic geometry, complex-analytic functions, and number theory.

To illustrate this connection, we discuss the *Markov Uniqueness*

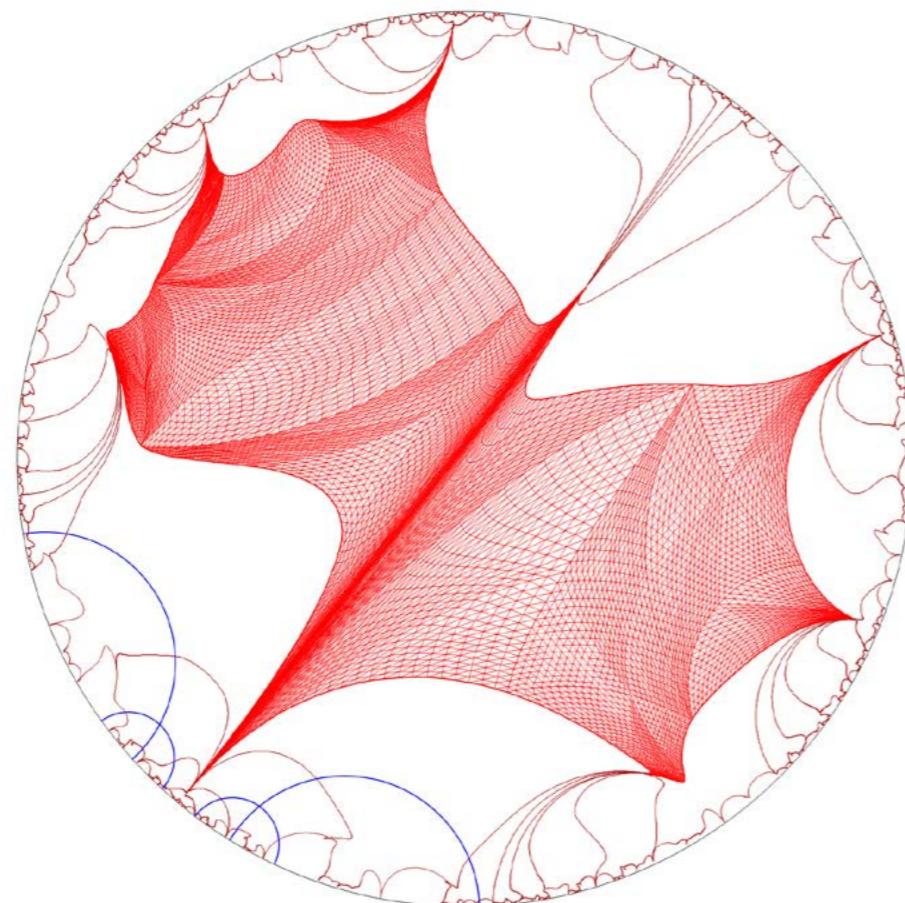


Figure 29: Equivariant harmonic maps of the hyperbolic plane in the Poincaré disk model (Pictures: Brice Loustau).

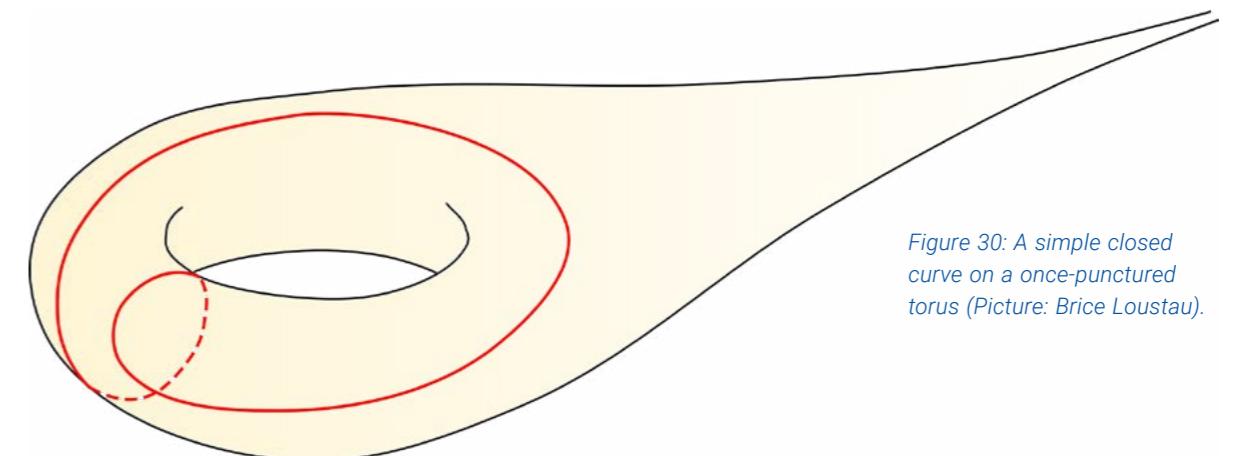


Figure 30: A simple closed curve on a once-punctured torus (Picture: Brice Loustau).

**Conjecture (MUC).** This problem concerns Markov numbers – that is, the integer solutions to the Diophantine equation  $x^2 + y^2 + z^2 = 3xyz$ , such as  $(x,y,z) = (1, 2, 5)$ ,  $(x,y,z) = (1, 5, 13)$ , etc. MUC asserts that there is exactly one such ordered triple  $(x, y, z)$  for each Markov number  $z$ . This notoriously difficult conjecture has defied solution by mathematicians for over a century.

It turns out that Markov numbers can be realized geometrically as the lengths of simple closed geodesics on a hyperbolic once-punctured torus known as the modular torus.

Using this beautiful connection, Gaster and Loustau [2020, The sum of Lagrange numbers, arXiv preprint: 2008.07659] proved that MUC is equivalent to the strikingly simple identity

$$\sum_{n=0}^{\infty} (3 - L_n) = 4 - \phi - \sqrt{2}.$$

In this formula,  $\phi$  is the golden ratio, and  $L_n$  stands for the well-known *Lagrange numbers*, which appear in the theory of Diophantine approximation and are related to Markov

numbers via the simple relation

$$L_n = \sqrt{9 - 4/M_n^2}$$

Numerical evidence strongly suggests that the identity holds and that MUC is therefore likely true.

## Minkowski spacetime and relativistic addition

In the early 20th century, the development and language of hyperbolic geometry was directly influenced by the rise of the theory of relativity in physics due to the relation of both theories to Minkowski spaces. Indeed, Minkowski's 4-dimensional "spacetime" is the mathematical framework for Einstein's theory of special relativity (in the vacuum). On the other hand, hyperbolic space (of any dimension) can be realized as the upper sheet of the hyperboloid of Minkowski space (of one dimension higher) consisting of unit timelike vectors (see Fig. 31, next page).

In this model of hyperbolic geometry, isometries are identified with orthochronous Lorentz transformations – that is, with time-orientation-pre-

serving isometries of Minkowski space. The hyperbolic analogs of Euclidean translations are identified with especially simple Lorentz transformations known as *Lorentz boosts*.

Einstein's theory of relativity is based on the postulate – confirmed via experimental measurements – that the speed of light in the vacuum is independent of the (freely falling) observer. It follows that the velocities of objects relative to moving frames cannot simply be added, as in Galilean physics. Instead, a velocity-addition formula exists that is well-known to physicists and that can be simply derived mathematically by composing a Lorentz boost and the projection of the hyperboloid on the  $t=1$  plane. This projection yields another famous model of hyperbolic geometry known as the Beltrami–Klein model (see Fig. 32, next page). The relativistic addition of velocities defines a "hyperbolic addition" in the Klein model of hyperbolic space analogous to the addition of vectors in Euclidean space.

## Hyperbolic geometry and machine learning

In recent years, interest in hyperbolic geometry has grown in the fields of data science and machine learning for representation learning via graph embeddings and for the construction of so-called hyperbolic neural networks. Learning graph representations via low-dimensional embeddings is an important problem in machine learning since many situations (e.g., in linguistics, evolutionary biology, computer networks, etc.) involve data that have a hierarchical structure, such as a graph structure. Hyperbolic spaces are well-known to geometers as typically being better ambient spaces for embeddings of graphs than are Euclidean spaces because the latter do not have "as much room" for the exponential growth of many graphs and trees. An example of the application of hyperbolic graph embeddings to historical-linguistics data is shown in Figure 33.

Hyperbolic neural networks were introduced by Ganea et al. [Hyperbolic Neural Networks, NeurIPS 2018], thereby connecting hyperbolic geometry with deep learning.

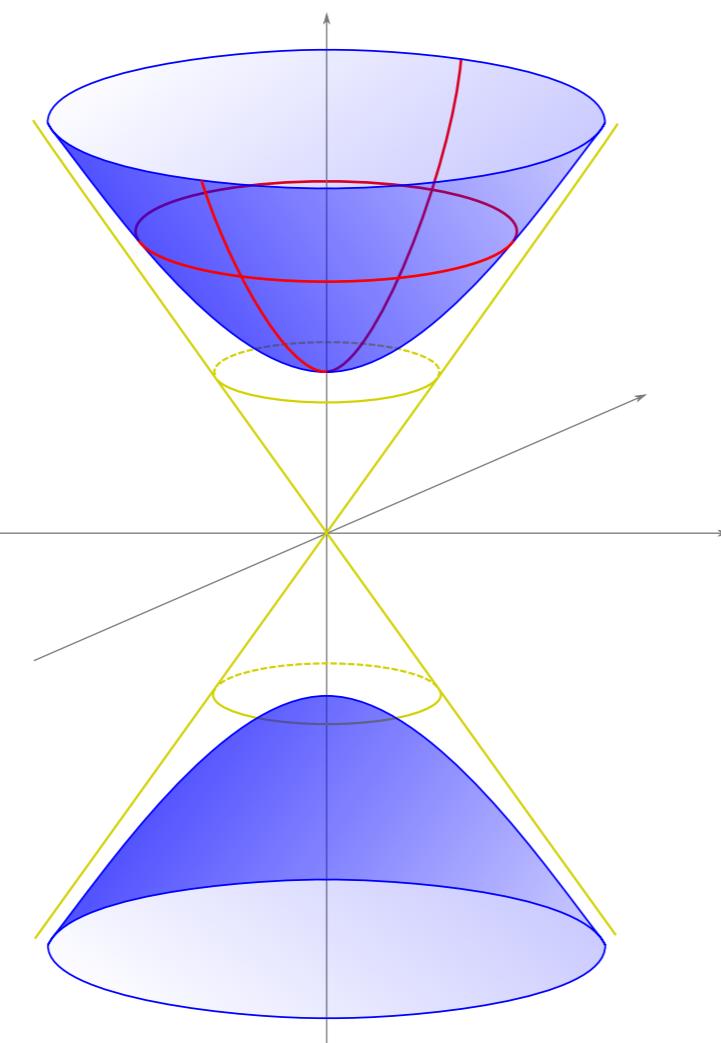


Figure 31: The hyperboloid in Minkowski space (Picture: Brice Loustau).

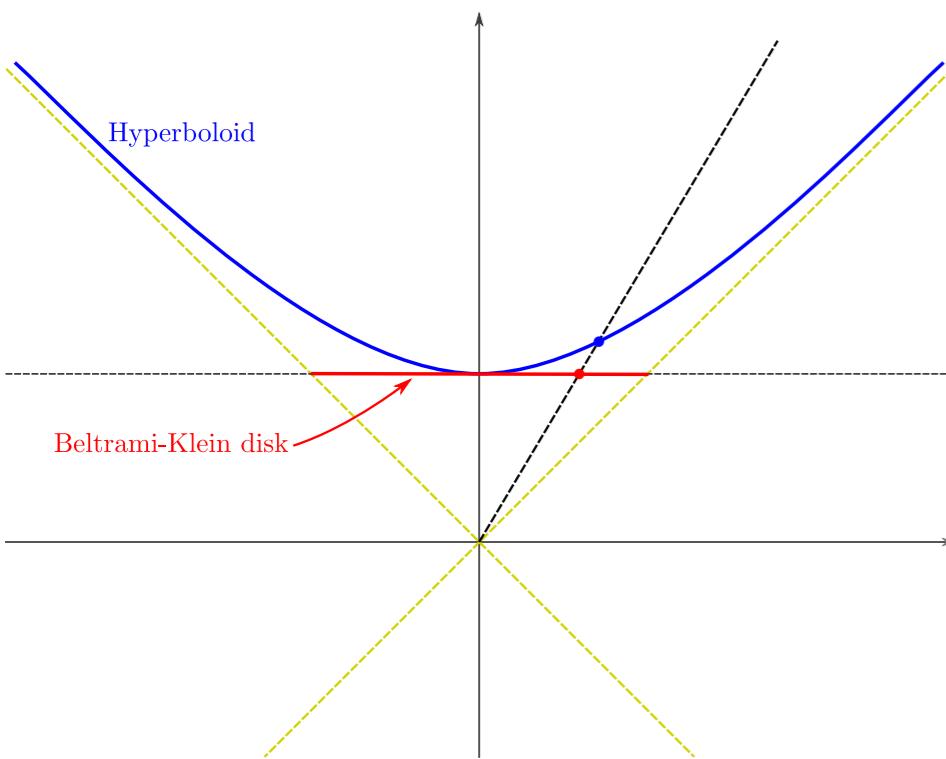


Figure 32: Schematic illustration of the projection of the hyperboloid onto the Beltrami-Klein model (Picture: Brice Loustau).

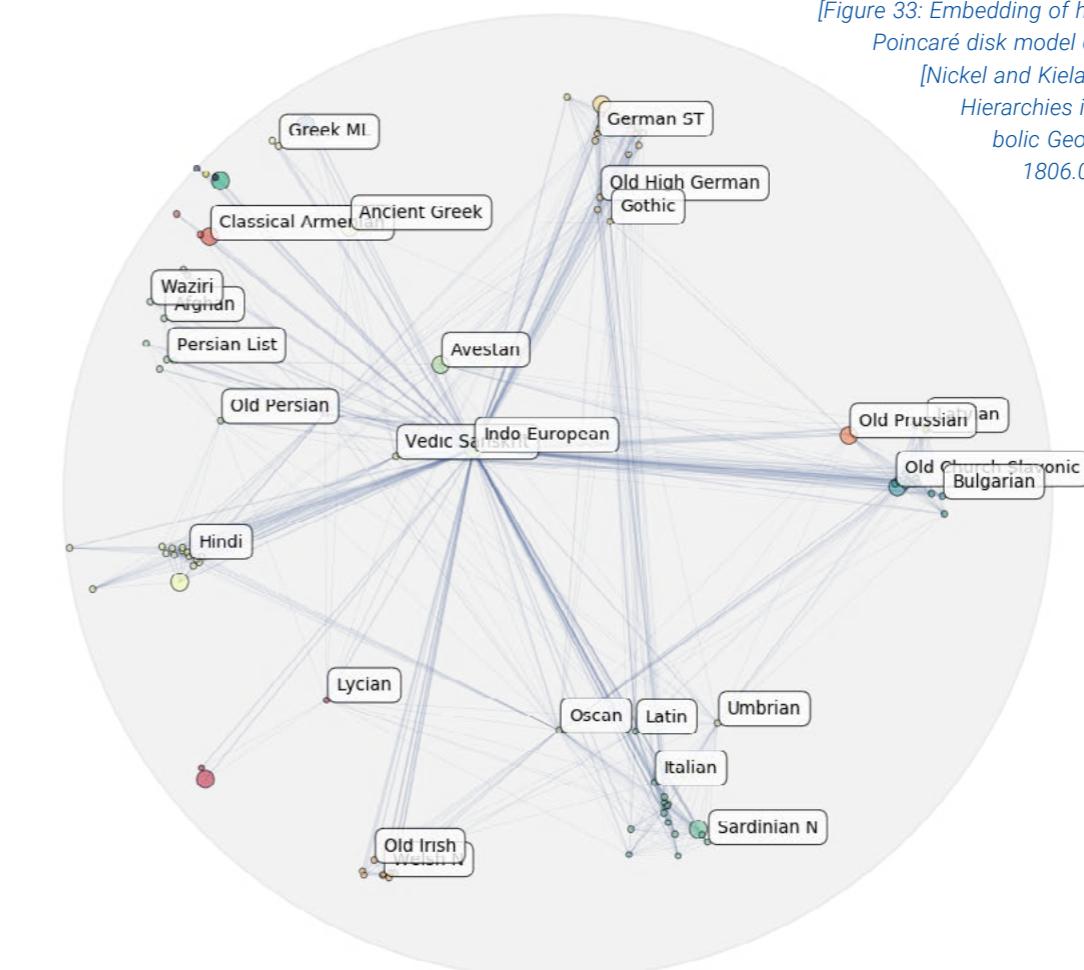
The key idea is to replace the linear operations that are used in standard neural networks with hyperbolic operations based on the "hyperbolic addition" described above. Empirically, hyperbolic neural networks appear well-suited for hyperbolic embeddings and have already been used or generalized by many authors. Members of the NLP group have applied hyperbolic geometry to fine-grain entity typing (see the two papers by [Lopez et al., 2020]). In 2020, we began a collaboration between the GRG and the NLP groups to explore the more-intricate non-Euclidean geometries that arise from symmetric spaces, which play an important role in the GRG group's research in the context of representation learning. A first article, in which we propose a systematic framework and metrics for learning graph embeddings in symmetric spaces, was submitted in February 2021.

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 dieses invariant lassen. Solche Transformationen lassen sich verknüpfen, d.h. hintereinander ausführen und bilden so die mathematische Struktur einer, so genannten, Gruppe. Im 19. Jh. entwickelte der Mathematiker Felix Klein einen neuen Begriff der Geometrie: Geometrie ist das Studium der Eigenschaften eines Raumes, die invariant sind unter einer 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 als auch die projektive Geometrie, die aus dem Studium der perspektivischen Kunst erwuchs und die nicht auf dem Messen von Abständen, sondern auf 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 **Groups and Geometry (GRG)** beschäftigt sich mit verschiedenen mathematischen Forschungsfragen auf dem Gebiet

der Geometrie, Topologie und der dynamischen Systeme, die das Zusammenspiel zwischen Räumen, wie zum Beispiel Mannigfaltigkeiten und metrische Räumen, und Gruppen, die als Symmetrien auf diese Räume wirken, einbeziehen. Außerdem wir beschäftigen uns mit den Anwendungen der Gruppentheorie und Geometrie in andere Disziplinen wie mathematische Physik, Datenwissenschaft und maschinelles Rechnen.



[Figure 33: Embedding of historical-linguistics data in the Poincaré disk model of the hyperbolic plane (from Nickel and Kiela, 2018, Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry, arXiv preprint: 1806.03417)].

Albanian
Anatolian
Armenian
Baltic
Celtic
Germanic
Greek
Indic
Indo-European
Iranian
Romance
Slavic
Tocharian

## 2 Research

### 2.7 Molecular Biomechanics (MBM)



#### Group Leader

Prof. Dr. Frauke Gräter

#### Staff members

Saber Boushehri (since June 2020)

Matthias Brosz

Svenja de Buhr

Florian Franz

Ana María Herrera-Rodríguez (until March 2020)

Dr. Fan Jin

Fabian Kutzki

Dr. Markus Kurth

#### Isabel Martin

Dr. Nicholas Michelarakis

Benedikt Rennekamp

Kai Riedmiller

Christopher Zapp

#### Students

Elizaveta Bobkova

Thomas Ehret (until December 2020)

Anna Schroeder (until March 2020)

Dennis Wagner

Simply speaking, the large variety of proteins in our bodies can be divided into two groups with vastly different jobs. The first group comprises enzymes, which take care of all biochemistry and whose functions range from building new molecules to degrading existing molecules and from attaching signals to silencing these signals. The second group comprises protein materials, which hold all parts of the body together, connect bones to muscles, surround cells with scaffolds, and protect the

cell nucleus and its precious genetic content from damage. In contrast to enzymes, protein materials are rather passive when it comes to biochemistry. They are made, modified, and degraded by enzymes but do not carry out any such sophisticated biochemistry themselves. Consider a virus: Its shell and adaptor proteins provide stability and establish connections, while the proteins contained on the inside of the virus capsid run the biochemical factory of an infection.

However, this view is currently being challenged, and the Molecular Biomechanics group was able to contribute a fresh perspective with our work on collagen in 2020. We found that rich biochemistry occurs within our supposedly passive fibers, networks, and scaffolds. Protein materials thus behave in this sense very much like an enzyme! We further outline these surprising findings below. The role of high-performance computations and

the development of new methods for better understanding such links between mechanics and biochemistry in proteins are critical and continuously growing, as exemplified in the following sections.

Below, we take a look back at a productive year of science full of social distancing and challenges to collaborative work and creative thinking.

#### Advances in molecular simulations of the mechanical properties and function of proteins

##### Florian Franz, Csaba Daday, and Frauke Gräter

Experimental observations and computer simulations are natural allies, as are single-molecule force spectroscopy (SMFS) and classical Molecular Dynamics (MD) when it comes to studying the response of molecules to mechanical force. Recent advances in experiments and simulations have increasingly facilitated a direct comparison of SMFS- and MD data, most importantly by closing the gap between timescales, which has traditionally been at least 5 orders of magnitudes wide. On the one hand, recent methodological advances in molecular simulations – primarily Molecular Dynamics (MD) simulations – have expanded the range of timescales that are accessible to simulations. On the other hand, emerging high-speed AFM methods allow for faster pulling speeds in experiments, which led to the recent breakthrough in the first direct quantitative comparisons of predicted rupture forces with experimentally measured forces. The basic principles of simulations under force have remained the same: An atom (or a group of atoms) in a molecule is (or are) subjected either to a virtual spring that is moved with constant velocity along the pulling direction or to a constant (or constantly changing) force while a counterforce or another

spring potential acts on another atom (or other atoms) of the molecular system to prevent the translation of the system. Significant progress has been made in three areas:

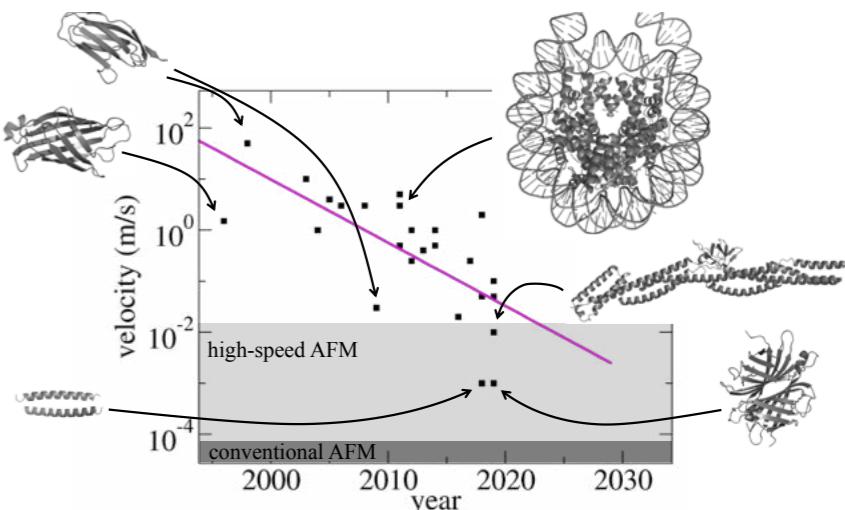


Figure 34. Pulling speeds used in MD simulations of proteins have decreased over time, approximately following Moore's law. Data points show the velocities of only a small subset of published simulations and include protein unfolding and protein-ligand-dissociation events with strongly varying system sizes and pulling lengths. The overview includes one data point from our group, namely from simulations of desmoplakin (structure on the right, center).

First, owing to the increased efficiency of hardware and MD codes, pulling velocities of far below 1m/s can now be reached in explicit solvent (Figure 34). Speeds as low as 0.001m/s have been reported in recent years.

Second, studies not only present single trajectories but can also include multiple or even dozens of trajectories for a given set of parameters. The large number of trajectories of recent studies has allowed for sufficient statistics on unfolding forces (under constant-velocity pulling) or on unfolding times (under constant-force pulling) as well as on

et al. How Fast Is Too Fast in Force-Probe Molecular Dynamics Simulations? The Journal of Physical Chemistry B, 123, 3658–3664, 2019).

Third, instead of focusing on the mechanical stability of individual protein domains, interest has recently shifted to larger multi-domain systems and to their mechanisms of mechano-sensing and mechano-transduction. Proteins in focal adhesions are particularly interesting candidates for such simulations of larger systems.

For the – admittedly only exemplary subset of – studies that Figure 34 is based on, pulling velocities have roughly halved every two years and

have thereby thus far followed Moore's law. However, on average, they have fallen behind the approximate doubling in protein MD simulation time every 1.3 years (Daday, C. et al.: The mechano-sensing role of the unique SH3 insertion in plakin domains revealed by Molecular Dynamics simulations, *Scientific Reports*, 7, 11669, 2017), which is likely due to the recent focus on improved statistics and larger biological systems discussed above. While Moore's law – which states that the number of transistors per area doubles every two years – is bound to be limited by physical constraints that will be reached within the coming years, we still expect the performance of simulations to increase drastically in the future. With GPU acceleration becoming more and more abundant and AI finding its way to simulations, the current trend is likely to even accelerate.

## Stretched beyond the limits

### Christopher Zapp

Polymers are subjected to mechanical stress in many everyday systems, such as those found in vehicle tires, shoe soles, and rubber bands. Imagine pulling a rubber band: Chemical bonds are cleaved and radicals form, leading to material damage. Because the yielding radicals are produced as a direct consequence of the mechanical force, they are called mechanoradicals. These mechanoradicals have been known for a century and have been technologically exploited. Now imagine your Achilles tendon. It is made of collagen, the virtual rubber band in your body. Collagen is the basic material of tendons, cartilage, ligaments, and bones that

provides structural and mechanical stability to nearly all human tissues. It is often under extreme pressure, for example, while you play sports or walk, during which time the Achilles tendon can experience multiple times your body's own weight in pressure. We asked ourselves: Do mechanoradicals also form in biological polymers – that is, in proteins such as collagen? Together with colleagues from Homburg, Frankfurt, and Seattle, we demonstrated in a series of experiments that excessive mechanical stress on collagen indeed produces radicals. The first challenge was to establish a setup to produce radicals inside collagen. A key experiment involved mounting and stretching a rat tail fascicle directly in an electron-paramagnetic-resonance (EPR) cavity in order to monitor radical formation in real time. By increasing the pulling force, the radical concentration inside the rat tail fascicle rose without any observed macroscopic ruptures. The 2-centimeter-long fascicle, which is mainly made of collagen I, survived forces of up to 20 Newtons for several minutes. The toughness of collagen comes from its rope-like micro-structure. A single collagen molecule is used to make larger collagen fibrils. The molecule is approximately 300 nanometers long and 1.5 nanometers in diameter and is made of three polypeptide chains. These chains or strands are twisted together into a right-handed triple helix. With type I collagen, the most-prominent type, each triple helix is called a collagen microfibril. These microfibrils are shifted relative to adjacent molecules by about 67 nanometers and build a representative unit of a collagen fibril. In each repeat unit of

the fibril, there is a part containing five molecules in cross-section, which is called the overlap region, and a part containing only four molecules, which is called the gap region. Covalent crosslinks between tropocollagen helices stabilize the fibril. Taken together, these unique and highly elaborated structural features of collagen equip it with high mechanical resilience. However, as the new experiments suggest, even if the collagen material remains intact at the tendon scale, individual chemical bonds inside the structure rupture, and radicals are formed. Molecular Dynamics simulations of the collagen fibril – which comprises millions of atoms – help to explain the experimental observations: Chemical bonds break at sites of high stress concentration when collagen is stretched. We built a collagen-fibril model in atomistic resolution to determine what happens during pulling at the microscopic level. The Molecular Dynamics simulations of collagen under constant force comprised a representative fragment of collagen, namely 27 triple helices at a length of one repeat unit along the fiber axis, which resulted in 5 million collagen- and water atoms, a challengingly large system to compute. By reaching a total of microseconds of simulation time, we were able to reveal the points of stress concentration and bond scission within the proteins. One major finding was that bonds within the crosslinks between triple helices as well as adjacent backbone bonds are particularly stressed and are candidates for bond scission. High-level quantum-chemical calculations revealed that the chemical nature of this scission is homolytic.

A striking feature of the collagen structure is the high abundance of aromatic residues (phenylalanines and tyrosines) in these regions. This observation was key to solving the riddle of EPR data: These aromatic residues served as first candidates for the radical signal, which was important in identifying the radicals. The resulting harmful radicals, which arise from covalent bond scission, are quickly scavenged by nearby aromatic residues – so-called DOPAs, or dihydroxyphenylalanines. This process prevents other side reactions and further degradation. The new discovery of a DOPA radical was supported by mass spectrometry, which revealed a specific DOPA side in one of the collagen chains. The DOPA radicals then finally convert into hydrogen peroxide, an important oxidative molecule in the body. Hydrogen peroxide in the concentrations we find in collagen acts as an important signaling molecule. Collagen is therefore not only a mere bearer of force, but it can also control the consequences of such forces. The study suggests

Vereinfacht gesagt kann man Proteine, wie sie in ihrer großen Vielfalt in unserem Körper vorkommen, in zwei Gruppen einteilen, die ganz unterschiedliche Aufgaben erfüllen: Die erste Gruppe umfasst die Enzyme. Sie kümmern sich um die gesamte Biochemie, vom Aufbau neuer Moleküle bis zum Abbau derselben, vom Anhängen von Signalen an andere Proteine bis zum späteren Ausschalten dieser Signale. Die zweite Gruppe von Proteinen umfasst die Baustoffe. Sie halten alle Teile des Körpers zusammen, verbinden Knochen mit Muskeln, umgeben Zellen mit Gerüsten und schützen den Zellkern mit seinem wertvollen genetischen Inhalt vor Beschädigungen. Im Gegensatz zu Enzymen sind diese Stoffe eher passiv, was die Biochemie angeht. Sie werden von Enzymen auf-, um- und abgebaut, führen aber selbst keine ausgefeilte Biochemie durch. Nehmen wir einen Virus: Seine Hülle und die Ankerproteine geben Stabilität und stellen Verbindungen her, während die im Inneren des Viruscapsids enthaltenen Proteine die biochemische Fabrik der Infektion betreiben. Diese Sichtweise wird derzeit in Frage gestellt, und die Gruppe **Molecular Biomechanics (MBM)** konnte mit ihrer Arbeit über Kollagen im Jahr 2020 zu einer neuen Perspektive beitragen. Wir haben herausgefunden, dass in unseren vermeintlich passiven Fasern, Netzwerken und Gerüsten eine reiche Biochemie stattfindet. Proteinmaterialien verhalten sich also in diesem Sinne sehr ähnlich wie ein Enzym! In diesem Bericht skizzieren diese überraschenden Erkenntnisse. Die Rolle von Hochleistungsrechnungen und Methodenentwicklung für das Verständnis solcher Zusammenhänge zwischen Mechanik und Biochemie in Proteinen steht außer Zweifel und wird stetig wichtiger. Wir blicken auf ein produktives Wissenschaftsjahr zurück, wenn auch mit physischer Distanz - mit all den damit verbundenen Herausforderungen für das gemeinsame Arbeiten und kreative Denken.

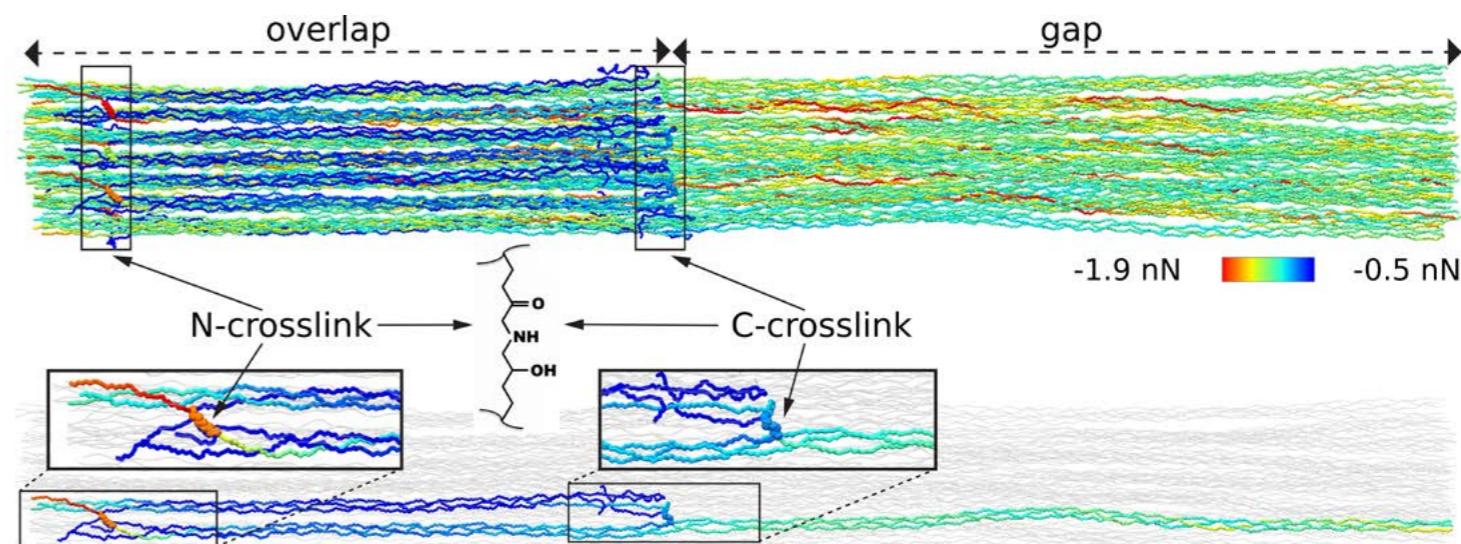


Figure 35. Force concentrates around crosslinks in tensed collagen. Snapshot of an MD simulation of the crosslinked collagen I model fibril under 1 nN of external force per chain, colored according to the distribution of the external force through the fibril (blue = low force, red = high force). Below is an example pair of overlapping triple helices connected by crosslinks from the snapshot, depicted separately to better visualize the forces around the crosslinks. The crosslinks are represented as spheres, the remaining collagen chains as gray ribbons.

## Where does collagen break?

### Benedikt Rennekamp

that collagen has evolved as a radical sponge to combat damage. In this way, collagen protects itself from the radicals (see Fig. 36). We are currently testing the hypothesis that collagen not only takes the mechanical load but acts directly as a biochemical-force sensor in our bodies. The coupling between mechanical and oxidative stress in collagen-based tissues is a missing link between the mechanical load and biological processes, from pain sensation to inflammation. These findings can explain why exercising can at times be quite painful and also provide a promising starting point for improving tissue repair and transplantation, for example, in sports medicine.

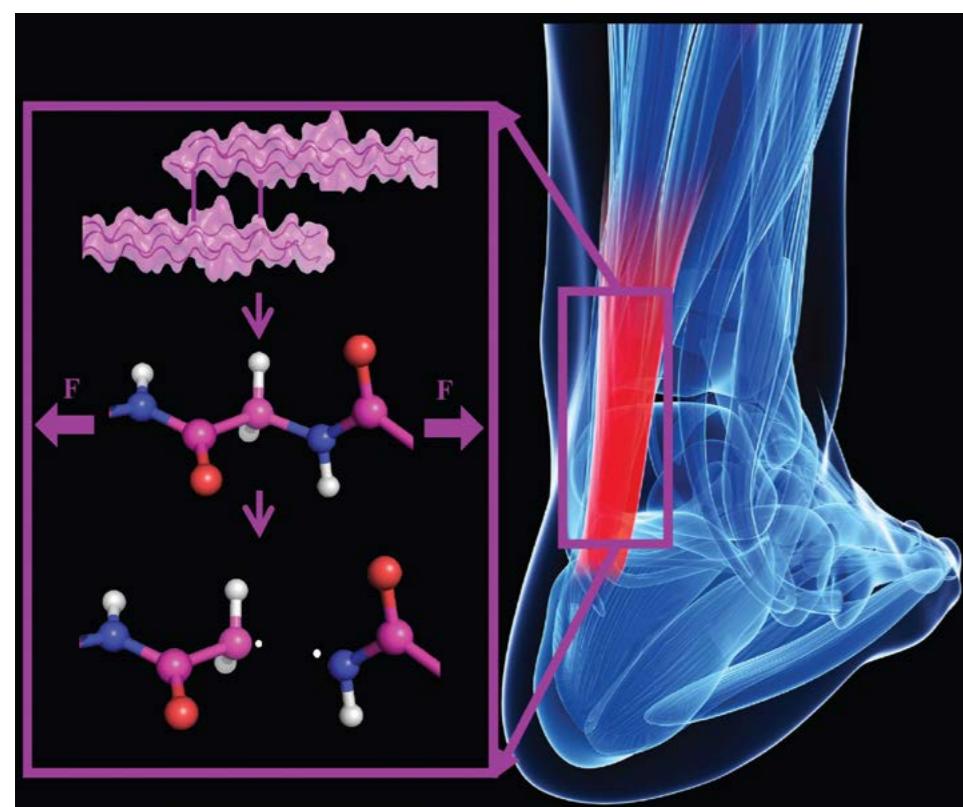


Figure 36. Forces in proteins, as in the displayed Achilles tendon, can lead to bond ruptures inside the molecule. These breakages, which can occur while the whole fibril is still intact, give rise to highly reactive radicals. (Photo credit - for the right part: istockphoto.com/SciePro).

To understand what happens at the molecular scale, Molecular Dynamics (MD) simulations are the standard tool that we employ in our

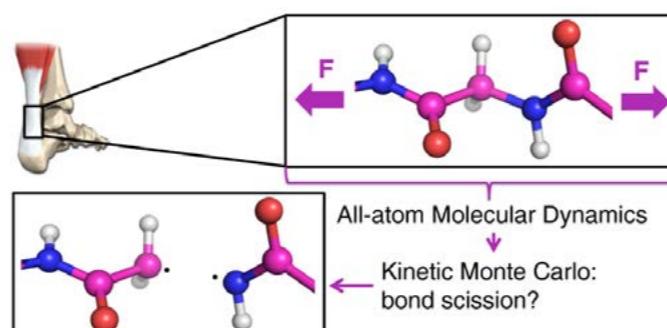


Figure 37. In order to simulate bond scissions in tensed proteins, we developed a hybrid kinetic Monte Carlo / Molecular Dynamics (KIMMDY) scheme that enables bond scissions in all-atom MD. simulations.

group. With respect to collagen, one question that arises in this context is how to identify weak spots that have a high rupture propensity.

However, in regular MD simulations, covalent bonds are predefined, and chemical reactions cannot occur. Furthermore, such events rarely take place at simulation timescales that are accessible with MD. Existing approaches that incorporate chemical reactions rely either on computationally expensive quantum calculations (e.g., QM/MM) or on complex bond-order formalisms in force fields (e.g., ReaxFF).

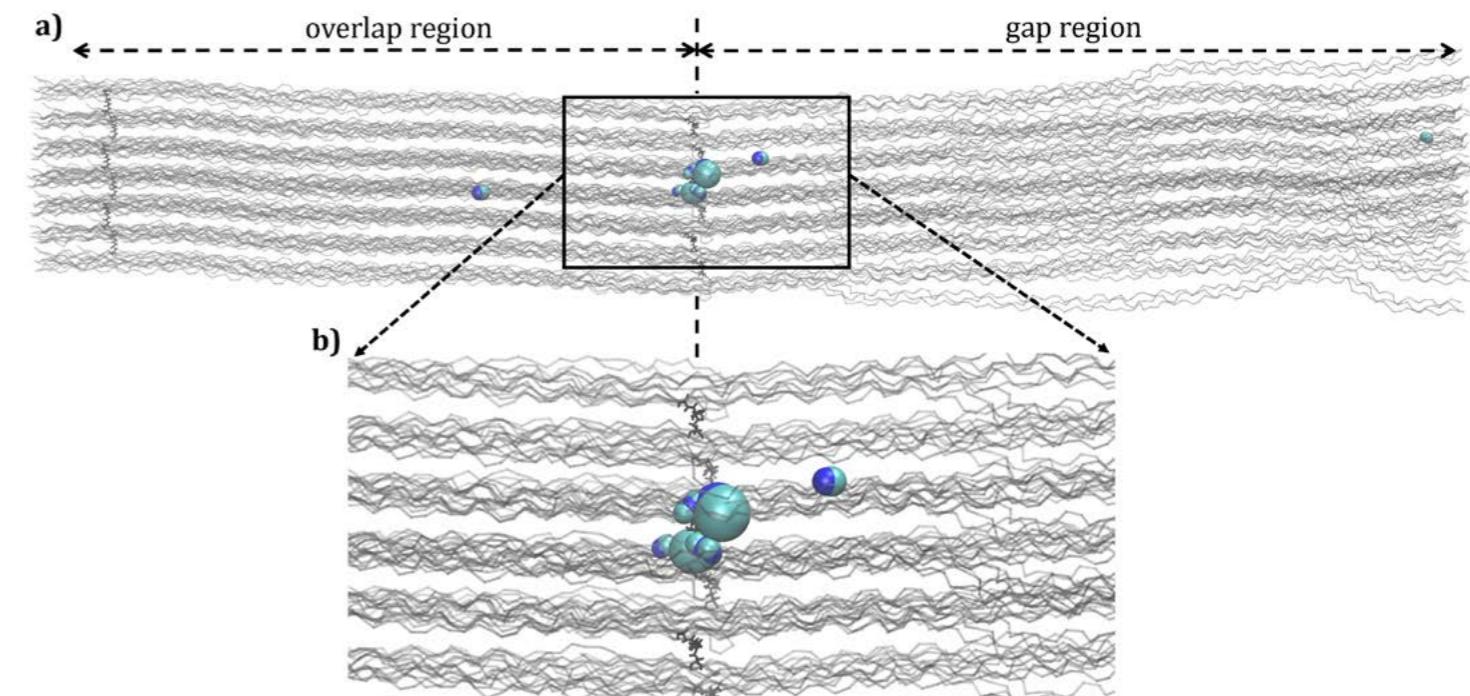


Figure 38. KIMMDY simulation of our previously modeled collagen fibril, consisting of one overlap and one gap region of the typical hierarchical collagen structure: At the transition of these regions, collagen strands that end are covalently connected to other molecules with crosslinks (shown here in black). The spheres mark simulated breakage sites that show a clear concentration in the vicinity of the crosslinks.

We therefore developed a new simulation scheme called KIMMDY (kinetic Monte Carlo / Molecular Dynamics) [citation of paper here] that can alleviate these issues via a hybrid combination of different simulation methods (see Fig. 37). Bond rupture rates are calculated based on the interatomic distances in the MD simulation and then serve as an input for a kinetic Monte Carlo step. This approach bridges the often-apparent separation of accessible timescales since the kinetic Monte Carlo step propagates the system in time until a rupture takes place. After this transition has occurred, KIMMDY creates a new molecular topology in order to switch

back to a Molecular Dynamics simulation. In this way, we are able to investigate the molecular response again after the bond scission and in greater detail. We applied this new technique to a multi-million-atom system of a tensed collagen fibril and showed that bond scissions clearly concentrate near the chemical crosslinks. These crosslinks connect different collagen molecules and are – according to our analysis – a potential weak spot in the collagen structure (see Fig. 38).

Switching back to the molecular simulations, the fibril can be seen to maintain its overall integrity after the

internal ruptures, even if the applied forces remain constant. These internal (homolytic) bond breakages lead to highly reactive radicals in collagen. The subsequently created species could potentially act in signaling processes by converting mechanical forces into oxidative stress. This method hence paves the way for further examinations of this biologically highly relevant mechanism at a molecular scale and complements our experiments in these fields (see the section "Stretched beyond the limits").

## 2 Research

### 2.8 Molecular and Cellular Modeling (MCM)



#### Group Leader

Prof. Dr. Rebecca Wade

#### Staff members

Christina Athanasiou

Manuel Glaser

Dr. Daria Kokh

Abraham Muniz Chicharro

Dr. Stefan Richter

Dr. Kashif Sadiq (until October 2020)

Alexandros Tsengenes

#### Visiting scientists

Madhura De (DKFZ)

Dr. Goutam Mukherjee (Heidelberg University)

Dr. Ariane Ferreira Nunes-Alves  
(Capes-Humboldt Fellowship)

Giulia Paiardi (EMBO Short-Term Fellowship, January–September 2020)

#### Students

Xingyi Cheng (January–March 2020)

Caroline Demidova (October–November 2020)

Jan-Niklas Dohrke (April–September 2020)

Sungho Bosco Han (since October 2020)

Anton Hanke (April–July 2020)

Lukas Jarosch (July–October 2020)

Konstantinos Mavridakis

Fabian Ormersbach (March–July 2020)

Jonathan Teuffel (August–November 2020)

Molecular recognition, binding, and catalysis are fundamental processes for cell function. The ability to understand how macromolecules interact with their binding partners and participate in complex cellular networks is critical 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 understand-

ing how biomolecules interact. What determines the specificity and selectivity of a drug–receptor interaction? How can proteins assemble to form a complex? 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? One of our aims is to gain a mechanistic molecu-

lar-level understanding of drug interactions along the process extending from drug delivery to drug–target binding to drug metabolism.

We take an interdisciplinary approach that entails collaboration with experimentalists and makes concerted use of computational approaches based on physics and bio-/chemo-informatics. The broad spectrum of techniques developed and employed ranges from interactive, web-based

visualization tools to machine-learning methods and atomic-detail molecular simulations.

In this report, we outline some of the results achieved in 2020.

Following a general overview of what was new in the group last year, we focus on projects dealing with (i) predicting drug–target binding kinetics, (ii) modeling macromolecular complexes, and (iii) structure-based drug design against SARS-CoV-2.

#### What was new in 2020?

It goes without saying that the coronavirus dominated our lives and work in 2020. In the MCM group, we not only adapted to working from home, to online teaching, to videoconference group meetings, and to virtual coffees, but we also sought to use our know-how to tackle the virus. We initiated several projects aimed at identifying therapeutics against the coronavirus, which are described below. In parallel, we continued work on our other projects. The final three-year phase of the EU-supported Human Brain Project began in April 2020. We participate in two collaborative tasks aimed at providing tools for using molecular-level information in multiscale simulations of signal propagation that will be incorporated into the EBRAINS infrastructure for digital brain research (<https://ebrains.eu>).

At the beginning of the year, Giulia Paiardi (University of Brescia, Italy) joined the group as a visiting PhD student supported by short-term fellowships from Erasmus+ and the EMBO. She performed simulations of glycoprotein interactions, including investigations of coronavirus spike-heparin interactions (see below). Having completed his Volkswagen Foundation “Experiment!” project on “RNA Epicatalysis,” Kashif Sadiq left the group in October to join the EMBL in Heidelberg. Ina Pöhner defended her doctorate on computational approaches to anti-parasite drug design against neglected tropical diseases in May, and after spending

some time in the SDBV group, she began work as a postdoc at the University of Eastern Finland, Kuopio, in January 2021. Jan-Niklas Dohrke completed his master’s thesis in Molecular Biotechnology, in which he performed simulations to investigate the mechanism behind the regulation of AMPAR ion channels in the brain by regulatory proteins. Fabian Ormersbach and Lukas Jarosch completed their bachelor’s theses in Molecular Biotechnology and Biochemistry, respectively, in the summer. Several master’s students from Heidelberg University completed internships in the group during the year: Xingyi (Cyan) Cheng (Molecular and Cellular Biology), Anton Hanke (Molecular Biotechnology), Jonathan Teuffel (Biochemistry), and Caroline Demidova (Chemistry).

The pandemic has changed the way scientific meetings are organized. Ariane Nunes-Alves co-organized two meetings that took place in September: a LatinXChem (<https://www.latinxchem.org/>) virtual conference, which was held exclusively on Twitter, and a virtual EMBL conference on “The impact of the COVID-19 crisis on women in science: Challenges and solutions” (<https://www.embl.org/events/covid19-wis>). Rebecca Wade was named lecturer for the 2020 Molecular Graphics and Modelling Society (MGMS) Lecture Tour (<https://www.mgms.org/WordPress/lecture-tour/>). Although the lectures, which had been planned to be held in five cities in the UK, had to be postponed due to the pandemic, Rebecca

did give an online lecture in November on “Computational Approaches to Protein Dynamics and Binding Kinetics for Drug Discovery”. While travelling came to a standstill, remote communication with scientists around the globe became easier, and activities aimed at strengthening the international scientific community grew in importance. Ariane Nunes-Alves’s activities as a member of the Early Career Board of the American Chemical Society’s Journal of Chemical Information and Modeling exemplify such activities. In 2020, Ariane published three articles on various aspects of conducting computational-chemistry research [Nunes-Alves, 2020b; Nunes-Alves, 2020c, Mazzolari, 2020]. She was also able to team up with MCM-group alumni Outi Sa-lo-Ahen (Turku University, Finland) and Ghulam Mustafa (DKFZ, Heidelberg) and with other scientists across Europe to write a review of the application of molecular-dynamics simulations for drug discovery and pharmaceutical development (Sa-lo-Ahen OMH, Alanko I, Bhadane R, Bonvin AMJJ, Honorato RV, Hossain S, Juffer AH, Kabeledev A, Lahtela-Kakkonen M, Larsen AS, Lescrinier E, Marimuthu P, Mirza MU, Mustafa G, Nunes-Alves A, Patsar T, Saadabadi A, Singaravelu K, Vanmeert M (2021).

Molecular dynamics simulations in drug discovery and pharmaceutical development.

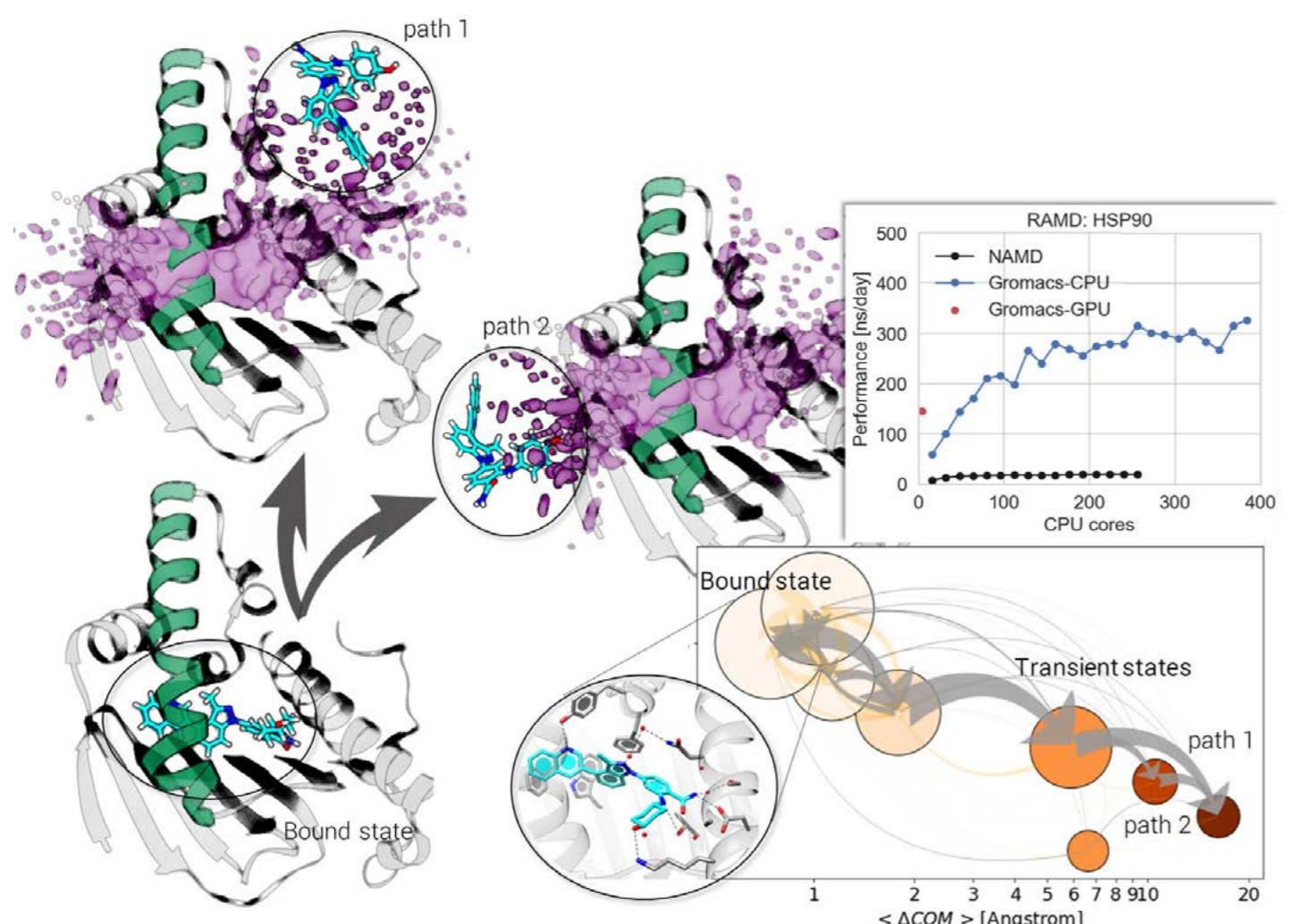
Processes 9(1):71.).

## Prediction of drug–target binding kinetics

Drug–target binding kinetics are increasingly understood to influence the efficacy of a drug. Therefore, drug-design procedures should aim to discover compounds that bind and unbind at optimal rates in addition to binding tightly to their molecular target. Consequently, there is a need

over a wide range of timescales. Recently, many new methods for computing binding kinetic parameters have been reported. We critically assessed their performance by considering two well-characterized benchmark systems [Nunes-Alves, 2020] and concluded that some of the methods can already be usefully applied for computing binding kinetic parameters in drug discovery lead

protein target – its residence time ( $\tau$ ) – is inversely related to the rate of dissociation of the drug–target complex. For most pharmaceutically relevant compounds, the timescales for dissociation from the target far exceed those that are accessible to conventional molecular dynamics simulations. To address this problem, we developed an efficient computational workflow that enables the

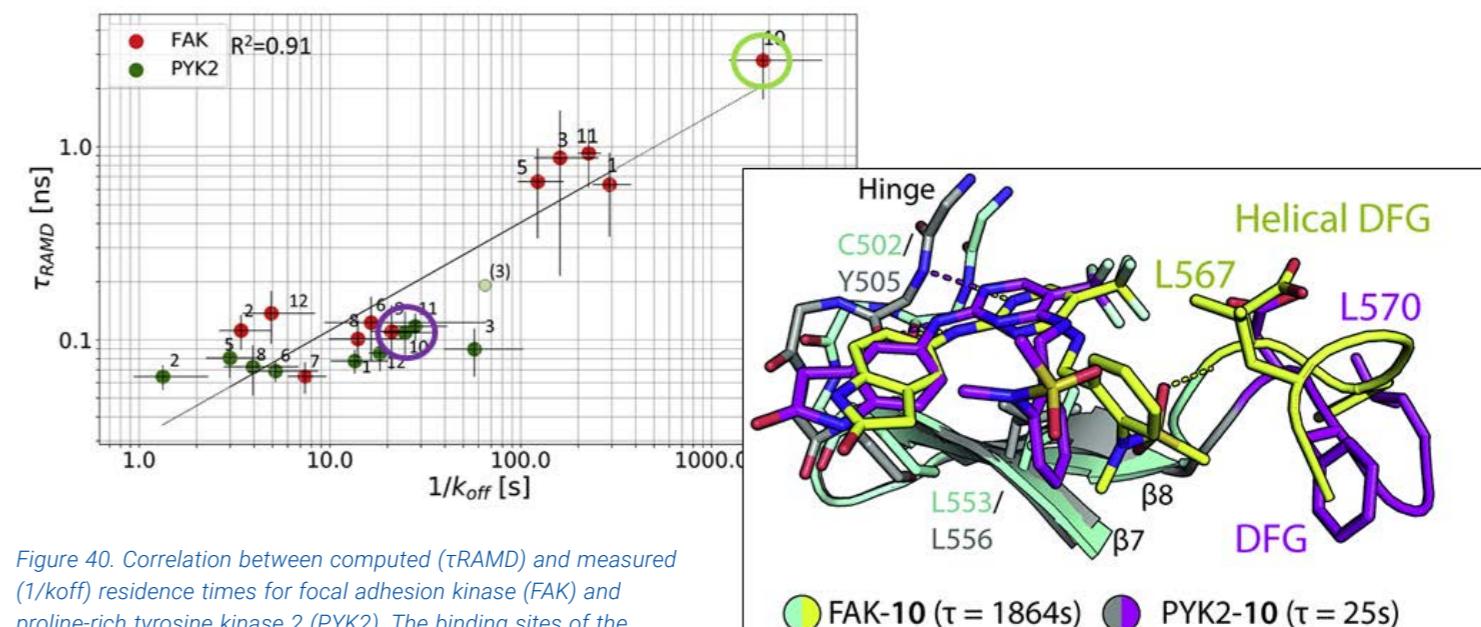


**Figure 39.** Illustration of the application of the  $\tau$ RAMD- and MD-IFP workflow for simulating the dissociation of a drug-like compound from an anticancer target protein. Purple contours indicate the regions explored by the compound during dissociation from the protein in RAMD simulations. The compound can dissociate via two main paths, along which different transient states, which influence the rate of dissociation, are sampled. The transient states are defined by an interaction fingerprint-based clustering of structures along the trajectories using MD-IFP. The RAMD simulations can be carried out efficiently using the implementation in the GROMACS software.

for accurate methods to compute binding kinetic parameters. However, computing drug–target binding kinetics poses a challenging problem because the rates at which drugs bind to and dissociate from proteins vary

optimization programs, but that further studies on more high-quality benchmark datasets are necessary to improve and validate computational methods. The length of time that a drug molecule spends bound to its

prediction of relative drug–protein residence times and the analysis of dissociation mechanisms in an automated manner. The workflow is based on simulations performed with the Random Acceleration Molecular



**Figure 40.** Correlation between computed ( $\tau$ RAMD) and measured ( $1/k_{off}$ ) residence times for focal adhesion kinase (FAK) and proline-rich tyrosine kinase 2 (PYK2). The binding sites of the complexes with compound 10, which has the longest residence time on FAK (green circle) and a comparatively low residence time on PYK (purple circle), are shown on the right. The overlay of the crystal structures of the complexes of compound 10 with FAK (cyan and yellow) and PYK2 (gray and magenta) reveals that the complex with FAK has a helical conformation of the DFG motif. FAK inhibitors with long residence times make hydrophobic interactions with the DFG motif in FAK and thereby induce a helical structure in FAK but not PYK2. (Figure adapted from: Berger et al.: Structure-kinetic relationship reveals the mechanism of selectivity of FAK inhibitors over PYK2, Cell Chemical Biology, 25 January 2021, with permission from Elsevier).

sity of Frankfurt), we combined experimental in vitro and in cellulo kinetic data, crystal structures, and mutagenesis data with  $\tau$ RAMD calculations of residence times

Dynamics (RAMD) method, which – in addition to existing implementations in the NAMD and AMBER software packages – Bernd Doser (Software Developer in the IT Services group) implemented in the freely available GROMACS molecular-simulation engine for simulations on CPU or GPU nodes, thereby enabling greatly improved computational performance (see Fig. 39). Relative dissociation rates are computed with our  $\tau$ RAMD protocol, and dissociation trajectories are analyzed using protein–ligand interaction fingerprints with our new MD-IFP set of tools [Kokh, 2020] (available at <https://github.com/HITS-MCM/MD-IFP>).

We apply the  $\tau$ RAMD- and MD-IFP workflow to a range of different proteins, including members of the important drug–target classes of kinases and G-coupled protein receptors. In one study performed with Stefan Knapp and colleagues (Univer-

to reveal the mechanism of kinetic selectivity of inhibitors between two closely related kinases that are targets for anticancer agents: focal adhesion kinase and proline-rich tyrosine kinase 2 (Berger B, Amaral M, Kokh DB, Nunes-Alves A, Musil D, Heinrich T, Schröder M, Neil R, Wang J, Navratilova I, Bomke J, Elkins JM, Müller S, Frech M, Wade RC, Knapp S: Structure-Kinetic-Relationship Reveals the Mechanism of Selectivity of FAK Inhibitors Over PYK2. Cell Chemical Biology, 25 January 2021, DOI: 10.1016/j.chembiol.2021.01.003 ; see Figure 40). The simulations enabled the prediction of relative residence times that were in good agreement with experimental data and the identification of the molecular determinants that affect the binding kinetics. They thus provide a basis for the rational optimization of compounds to extend residence times. This study demonstrated how small differences between protein sequences

can affect the interplay between protein structural mobility and ligand-induced structuring, thereby leading to the kinetic selectivity of enzyme inhibitors.

## Modeling of macromolecular complexes

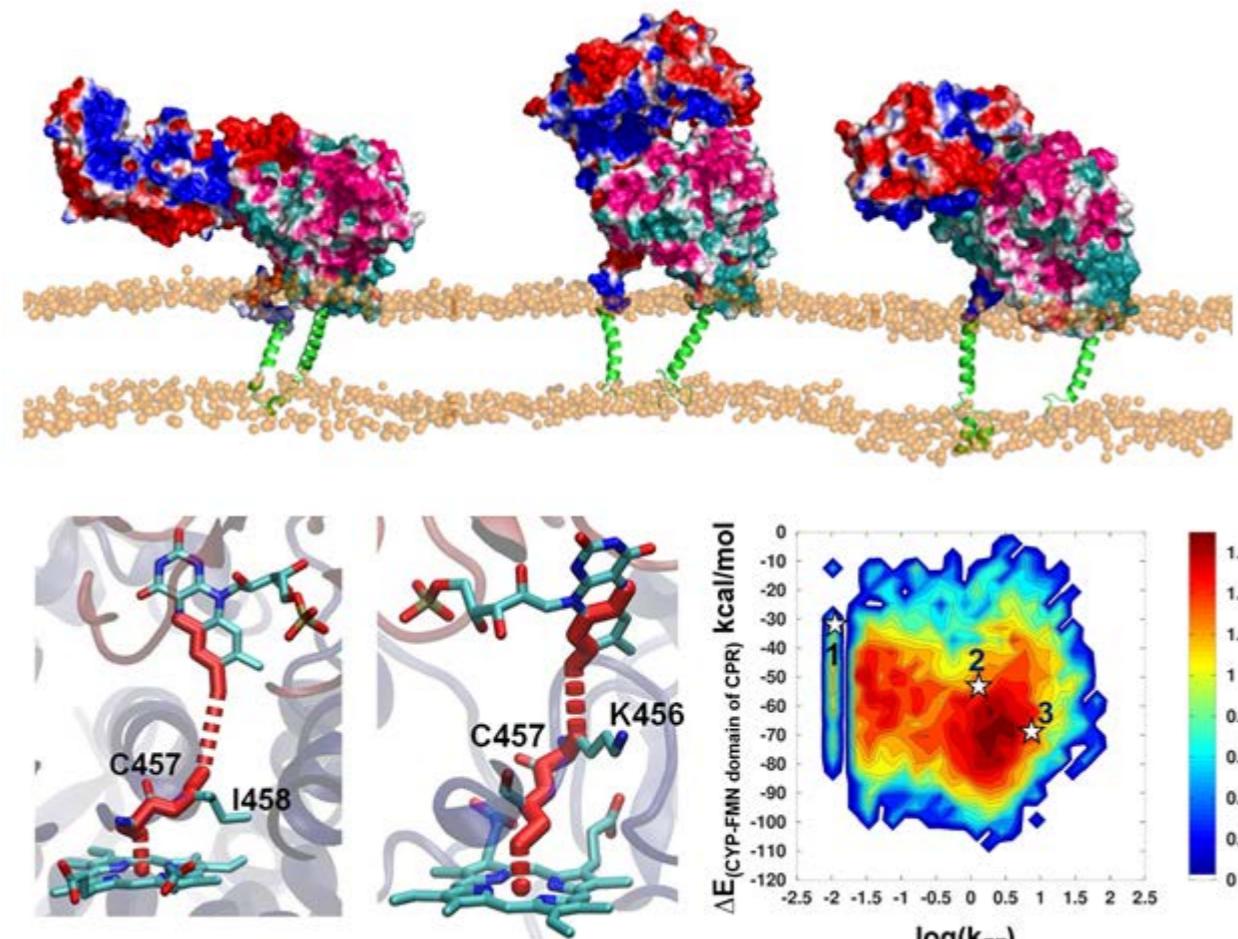
We employ structural bioinformatics and molecular simulation approaches in combination with experimental data to predict structures of protein–protein– [Diestelkoetter-Bachert, 2020, Moreau, 2020/2021], protein–peptide– [Weidner, 2020], and protein–nucleic-acid [Öztürk, 2020], [Öztürk, 2020b] complexes and to investigate binding mechanisms and properties. The modeling and simulation of protein complexes in membrane environments presents particular computational challenges, which we address in several projects, including the Euro-neurotrophin ITN network (<https://www.euronurotrophin.eu/>) and the Informatics4Life consortium (<https://informatics4life.org/>). In this report, we describe a multiresolution simulation approach for investigating how membrane-bound mammalian cytochrome P450 (CYP) enzymes form electron transfer-competent complexes with their redox partner, cytochrome P450 reductase (CPR).

CYPs form a superfamily of ubiquitous heme-containing monooxygenases that catalyze drug metabolism and steroidogenesis and are of significant interest for exploitation as biocatalysts and drug targets. The CYP catalytic cycle requires two electrons to be transferred to the heme cofactor, and the electron-transfer steps have been observed to be rate-limiting for the reaction in many cases.

Both CYP and CPR are anchored in the membrane of the endoplasmic reticulum by a single transmembrane helix. The globular catalytic domain of CYPs is attached to the transmembrane helix by a flexible linker. We combined coarse-grained and atomic-detail molecular dynamics simulations to predict how

CYPs arrange in a phospholipid bilayer with the catalytic domain dipping into the membrane [Mustafa, 2020]. We used such a model of CYP 1A1, which

plays an important role in the metabolism of carcinogens, such as those in tobacco smoke, as a starting point for modeling CYP1A1-CPR complexes. We first performed rigid-body Brownian dynamics docking simulations of the CYP catalytic domain with the N-ter-



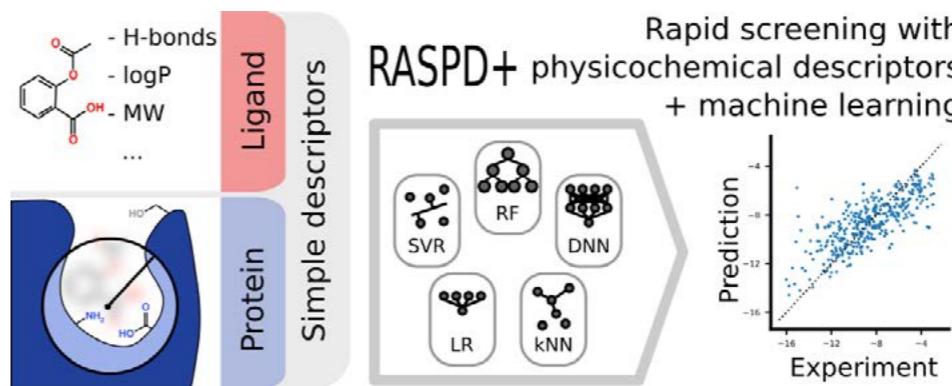
**Figure 41.** Representative structures and predicted electron transfer pathways for the structural ensemble of electron-transfer-competent CYP 1A1–CPR complexes in a phospholipid bilayer obtained via a multiresolution dynamics simulation approach. Upper panel: Three complexes shown with protein surfaces colored by electrostatic potential (positive: CYP (cyan) and CPR (blue); negative: CYP (pink) and CPR (red)), with transmembrane helices in green and the lipid phosphorous atoms displayed as orange spheres. Lower panel: The two predicted electron-transfer pathways from the CPR FMN cofactor to the CYP heme cofactor in complexes are indicated in red in close-up views of the protein–protein interfaces. The 2D-histogram plot shows the computed interaction free energy between the CYP globular domain and the CPR FMN domain for snapshots from the simulations versus the computed electron transfer rates. Stars indicate the complexes shown in the upper panel. Adapted from: [Mukherjee, G., Nandekar, P.P. & Wade, R.C. An electron transfer competent structural ensemble of membrane-bound cytochrome P450 1A1 and cytochrome P450 oxidoreductase. *Commun Biol* 4, 55 (2021). <https://doi.org/10.1038/s42003-020-01568-y> (CCBY 4.0)].

Stuttgart. We found that upon binding to CPR, the CYP 1A1 catalytic domain becomes less embedded in the membrane and reorients, indicating that CPR may affect substrate passage to the CYP active site from the membrane. Despite the constraints

imposed by membrane binding, we identified several arrangements of CPR around CYP 1A1 that are compatible with electron transfer (see Figure 41). Computed electron-transfer rates and pathways agree well with available experimental data and provide evidence as to why the CYP–CPR electron-transfer rates are low compared with those of soluble bacterial CYPs. Moreover, the identified elec-

## Structure-based drug design against SARS-CoV-2

During 2020, we pursued three strategies toward antiviral therapeutics against SARS-CoV-2.



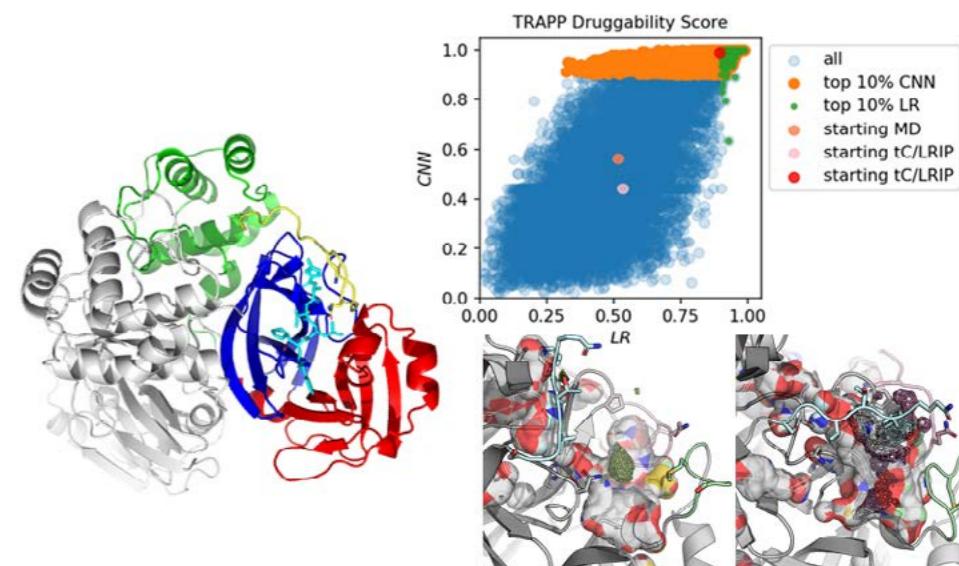
**Figure 42.** Schematic overview of the workflow of RASPD+ (Rapid Screening with Physicochemical Descriptors + machine learning) from [Holderbach, 2020]. For each ligand molecule, simple physicochemical descriptors are computed based on atomic contributions. Information on the target protein is gathered within a sphere around a putative binding position. A set of machine-learning (linear regression (LR), k-nearest neighbors (kNN), support vector regression (SVR), neural network (DNN), and random forest (RF)) methods is used to train and validate models, which are then used to predict binding free energies.

**(1)** We pooled the expertise of the MCM group together with that of Francesca Spyrikis (University of Turin) and Giulia Rossetti (Forschungszentrum Jülich) in order to participate in the “JEDI billion molecules against Covid19 grand challenge” (<https://www.jedi.foundation/billion-molecules>) and to virtually screen millions of compounds for the inhibition of four virus proteins. Along with about 130 other teams worldwide, we submitted prioritized lists of compounds based on the results from our computational pipeline. JEDI analyzed all submissions and selected 1,000 compounds for synthesis and experimental testing, including compounds on the lists that we submitted. Tests of the compounds remain ongoing.

This work required a very large number of compounds to be screened, which we achieved using RASPD+ (Rapid Screening with Physicochemical Descriptors + machine learning) [Holderbach, 2020]. We developed RASPD+ as a fast pre-filtering method for ligand prioriti-

zation that is based on a set of machine-learning models and uses simple pose-invariant physicochemical descriptors of the ligands and the target protein-binding pocket to estimate protein–ligand binding

**(2)** In virtual screens, the docking of compounds is typically performed to only one or a few structures of the target protein. However, proteins are flexible, and their binding sites can change shape and reveal transient or cryptic pockets that can provide binding sites for compounds. We developed TRAPP (TRAnsient Pockets in Proteins; [trapp.h-its.org](http://trapp.h-its.org)) to enable the exploration of different protein conformations, the analysis of binding pocket flexibility and dynamics, and the extraction of spatial and physicochemical information on the binding pocket conformations. Version 4 of the TRAPP webserver, which includes a machine learning method to score pocket druggability [Yuan, 2020], was released this year. We applied TRAPP to analyze the binding pocket dynamics of the SARS-CoV-2 main protease and the spike receptor-binding domain. For the main protease, we analyzed about 30,000 conformations of the protein – from crystal structures, molecular dynamics simulations, and conformations generated by TRAPP – to select structures with high druggability scores (see Figure 43).



**Figure 43.** Analysis of the flexibility and druggability of the SARS-CoV-2 main protease showed high flexibility of the active site and revealed that small structural variations dramatically impact ligand-binding properties. Left: A structure of the main protease homodimer with an inhibitor (cyan) bound in the active site. Upper right: Distribution of the two TRAPP druggability scores for computationally generated structures. Lower right: Binding sites of two representative structures with high druggability scores, with pocket surfaces colored by atom type. The high mobility of the flexible loops (green and pink) and the C-terminal tail (cyan) should be noted. Adapted from [Gossen et al., 2021].

These conformations were used for virtual screening against compound libraries by Giulia Rossetti and her colleagues (Forschungszentrum Jülich), who were able to identify a blueprint for a specific structure-based pharmacophore, which was used to predict high-affinity inhibitors that were independently validated both in an enzyme-inhibition assay and by crystallography (see: Gossen J, Albani S, Hanke A, Joseph BP, Bergh C, Kuzikov M, Costanzi E, Manelfi C, Storici P, Gribon P, Beccari AR, Talarico C, Spyros F, Lindahl E, Zaliani A, Carloni P, Wade RC, Musiani F, Kokh DB, Rossetti G: A blueprint for high affinity SARS-CoV-2 Mpro inhibitors from activity-based compound library screening guided by analysis of protein dynamics. ACS Pharmacology & Translational Science. Xx, 2021).

**(3)** In collaboration with Marc Rusnati and Giulia Paiardi (University of Brescia, Italy), we performed molecular dynamics simulations to investigate the interactions of heparin with the SARS-CoV-2 spike glycoprotein. Heparin is polysaccharide administered intravenously as an anticoagulant to COVID-19 patients and via aerosol for treating other lung diseases. Experiments indicate that heparin inhibits SARS-CoV-2 infection by binding to the virus spike glycoprotein, which plays a key role in attachment and fusion with host cells by interacting with the host-cell ACE2 receptor and heparan sulfate proteoglycans, which are structural analogues of heparin that act as co-receptors and may influence host susceptibility. Our molecular dynamics simulations were designed to investigate the role of heparan sulfate proteoglycans in SARS-CoV-2 infection and the inhibitory activity of heparin. We modeled the homotrimeric head of the spike glycoprotein in active and inactive

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 strukturbasierte Wirkstoffentwicklung.

In der Gruppe **Molecular and Cellular Modeling (MCM)** 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 begrenzte Zellumgebung auf die Bildung eines Proteinkomplexes? Warum verlaufen einige Bindungsprozesse schnell und andere langsam? Welche Auswirkungen haben Proteinbewegungen auf ihre Bindungseigenschaften?

Eines unserer Ziele besteht darin, die Mechanismen besser zu verstehen, die bei Wechselwirkung von Medikamenten auf der molekularen Ebene ablaufen, von der Freisetzung des Wirkstoffs über die Bindung zum Rezeptor bis hin zum Metabolismus des Medikaments.

In einem interdisziplinären Ansatz kooperieren wir mit experimentell arbeitenden Forscher/-innen und verwenden gemeinsam rechnerische Methoden aus den Bereichen der Physik-, Bio- und Chemoinformatik. Das breite Spektrum der Techniken, die wir entwickeln und einsetzen, reicht dabei von interaktiven web-basierten Visualisierungswerkzeugen bis hin zu Molekulsimulationen auf atomarer Ebene.

In diesem Bericht beschreiben wir einige der Ergebnisse aus dem Jahr 2020. Nach einem allgemeinen Überblick über Neuigkeiten in der Gruppe konzentriert sich der Bericht auf Projekte, die sich mit (i) der Vorhersage von Wirkstoff-Protein-Bindungskinetik, (ii) mit der Modellierung von makromolekularen Komplexen und (iii) mit strukturbasierter Wirkstoffentwicklung gegen SARS-CoV-2 befassen.

prefusion conformations with zero, one, or three bound heparin oligosaccharides. We then performed several replica microsecond simulations of each system. Due to the large size of these systems (about 700,000 atoms), these simulations were run on the PRACE Marconi100 GPU nodes at CINECA, Italy. Our models reveal long, positively charged patches on the spike head that can accommodate the linear anionic polysaccharide chains of heparin or heparan sulfate proteoglycans. On binding at these patches,

heparin masks key functional sites on the spike, stabilizes it in its closed inactive conformation, and allosterically modulates the exposure of the host-receptor binding site in the open, active conformation. Our results provide a basis for the rational optimization of heparin derivatives for antiviral therapy.

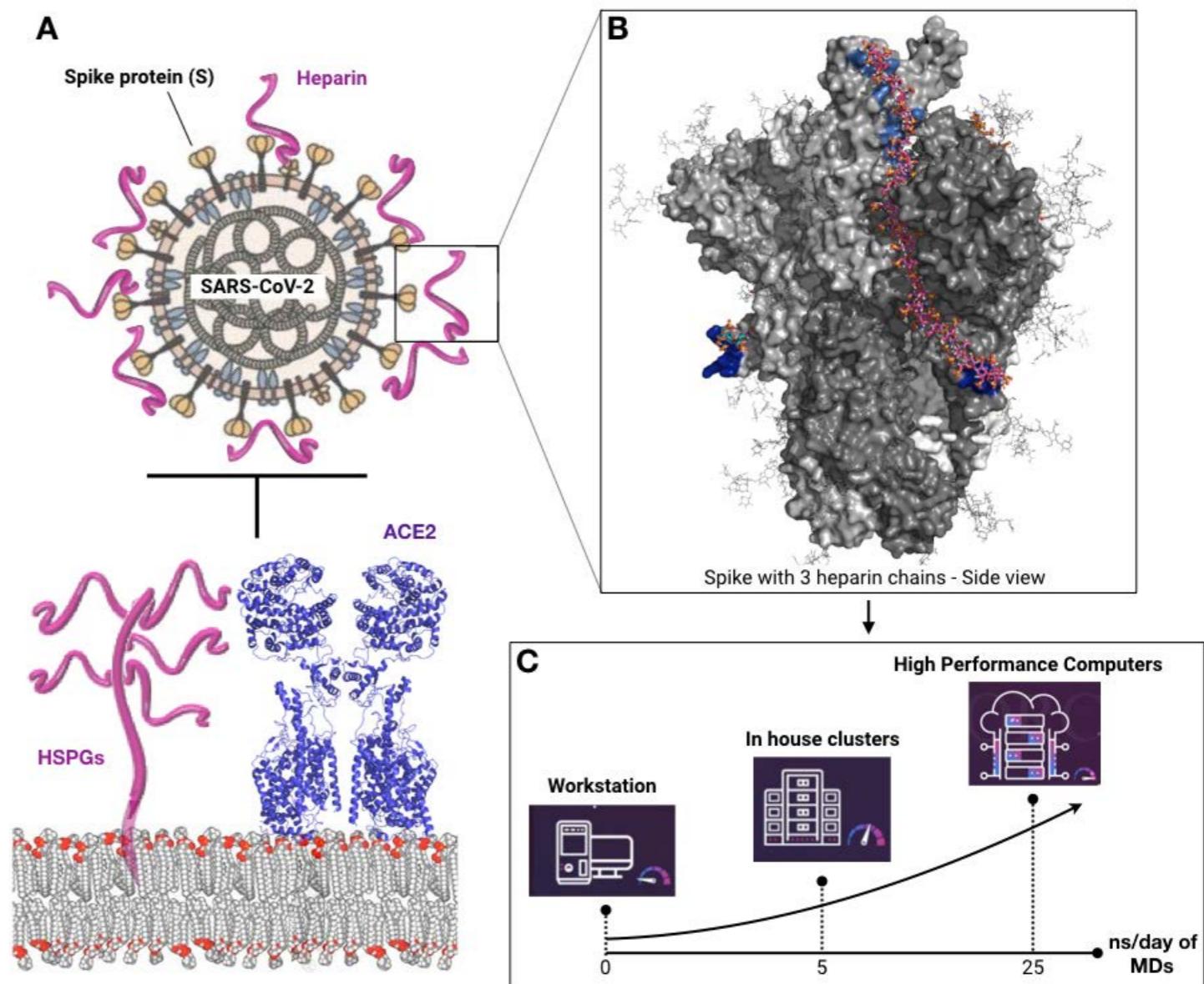
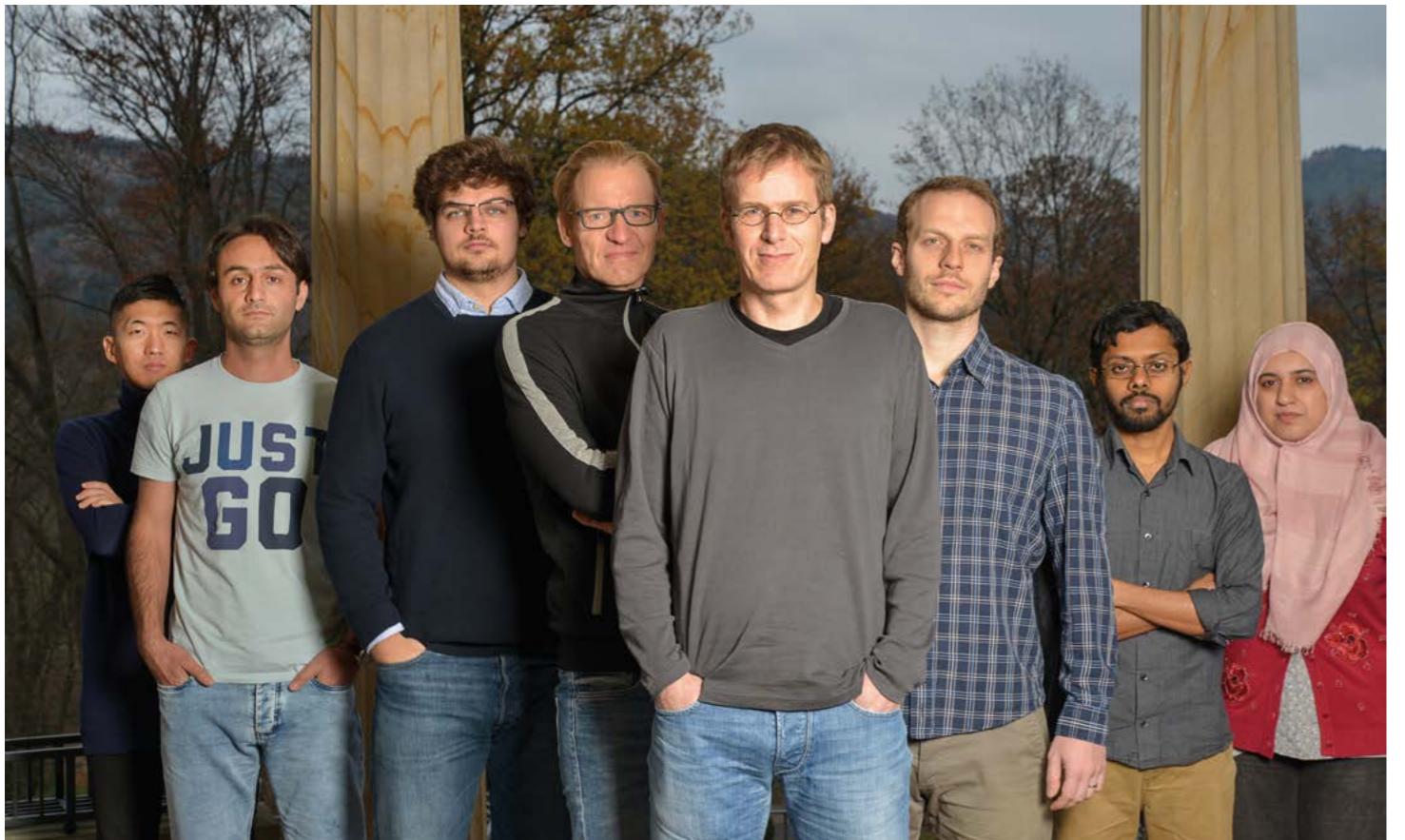


Figure 44. Schematic illustration of our molecular dynamics simulation approach to studying interactions of the SARS-CoV-2 spike with heparin. (A) Recent findings suggest that SARS-CoV-2 infection depends on the interaction of the SARS-CoV-2 spike both with the glycan chains of the heparan sulphate proteoglycans (HSPGs) and with the angiotensin-converting enzyme-2 receptor (ACE2) of the human host cell and that heparin might interfere with these interactions. (B) View of the atomic-detail model of one of the simulated systems consisting of the homotrimeric spike glycoprotein head (gray) – with one subunit in an open conformation suitable for binding ACE2 – interacting with 3 heparin chains (pink), each composed of 31 monosaccharides. (C) Use of PRACE high-performance computers enables such systems to be simulated about 25 times faster than on a workstation and 5 times faster than on a typical in-house compute-cluster. Image published in "PRACE Digest 2020" (extract in: <https://prace-ri.eu/interactions-of-the-spike-protein-and-heparin>).

# 2 Research

## 2.9 Natural Language Processing (NLP)



### Group Leader

Prof. Dr. Michael Strube

### Staff member

Dr. Mark-Christoph Müller

### Scholarship holders

Haixia Chai (HITS Scholarship)

Kevin Alex Mathews (HITS Scholarship)

Sungho Jeon (HITS Scholarship)

### Visiting scientists

Mehwish Fatima (PhD student, HEC-DAAD Scholarship)

Federico López (PhD student, Research Training Group AIPHES, delegated, Heidelberg University)

### Students

Jason Brockmeyer

Fabian Düker

Lucas Rettenmeier (until June 2020)

Tobias Martiné

Natural Language Processing (NLP) is an interdisciplinary research area that lies at the intersection of computer science and linguistics. The NLP group develops methods, algorithms, and tools for automatically analyzing natural language. The group focuses on discourse processing and related applications, such as automatic summarization and readability assessment.

machine-learning methods – such as support-vector machines – to neural networks. In his last published work at HITS, he created a neural network inspired by linguistic theories. Mohsen currently works as a post-doctoral researcher at the UKP lab at the Technical University of Darmstadt. Master's student Lucas Rettenmeier, who joined the NLP group in 2019, submitted his master's thesis at Heidelberg University in July 2020 – interestingly, the thesis was submitted to the Department of Physics and Astronomy. Lucas worked on the stability of word embeddings and applied his insights to determine semantic change over time. He has since joined Amazon Web Services as a Solutions Architect. Along the way, Lucas also finished the half marathon at the (virtual) NCT Run 2020 in less than one hour and 20 minutes (see Chapter 4)!

The NLP group continued to publish in 2020 at first-rate conferences – such as EMNLP and COLING – with first-authored papers by Sungho Jeon and Federico López. Over the summer, Federico completed an internship at Google Research in Mountain View, California (remote internship).

The NLP group also began two new collaborations with other research groups at HITS in 2020. Together with the SDBV group, we initiated DeepCurate, a BMBF-funded project with the aim of (semi-)automatically curating database entries in the biomedical domain. Another exciting collaboration developed around Federico López's work on hyperbolic neural networks for entity linking. This work caught the attention of the GRG group and led to a joint workshop publication by Federico, Beatrice Pozzetti, Steve Trettel (Stanford University), and Anna Wienhard (see Chapter 2.6, pp. 44/45). We continue to work on geometric deep learning and will further this work in a HITS lab project in 2021.

Michael Strube was co-chair of the "First Workshop on Computational Approaches to Discourse," which took place (virtually) at EMNLP 2020. The event comprised invited talks, paper presentations, and a large panel discussion on creating a community of discourse researchers in NLP. The second iteration of the workshop is already in the works and will hopefully be held in person at the EMNLP 2021.

### Neural-network-based coherence modeling

#### Sungho Jeon

Coherence describes the semantic relationship between the elements of a text. A text passage can be analyzed either as a unified whole or as a collection of unrelated sentences. The most well-known computational theory – centering theory – was proposed in 1983 by Grosz, Joshi, and Weinstein. Centering theory determines the most-salient item in each sentence – the center (or focus) – and tracks changes in this center. Designing an AI system inspired by this theory in 2020 required overcoming many challenges.

One challenge to the design involved how to capture the focus of each sentence. Prior AI studies on coherence have simplified the task and

mostly used entity information to describe the focus. These studies assume that entities are already determined and track focus changes by checking which entities occur in sentences and whether these entities are identical. This approach renders centering theory accessible to AI systems, and many studies follow this idea. However, these systems are mostly based on heuristics and do not realize the core elements of centering theory. In 2020, we published two papers at the EMNLP [Jeon and Strube, 2020a] and COLING conferences [Jeon and Strube, 2020b] that aimed to combine a linguistically thorough treatment of centering with modern neural-network-based AI techniques.

Centering only considers changes to the focus at the sentence level. Coherence, however, arises not only at the sentence level but also at the document level, thereby providing insights into the structure of the whole text or discourse. Discourse structure represents the semantic organization of a whole text. Incorporating structural information into the centering model is beneficial for diverse downstream tasks. In order to identify the discourse structure, earlier works have adopted a supervised approach that relies on human linguistic annotation. However, annotating the discourse structure is difficult, time-consuming, and costly and requires that annotators understand not only the local context

surrounding the target sentence but also the higher discourse-level relations.

In [Jeon and Strube, 2020a], we propose a coherence model inspired by centering theory that accounts for structural information. The model does not rely on human annotations to identify this information and consists of two components: (1) a discourse-segment parser that determines structural relationships between discourse segments by tracking focus changes between segments and (2) a structure-aware transformer that exploits structural information to update sentence representations. The discourse-segment parser first identifies the hierarchical discourse segments of a text, thereby building on an approximation of centering theory. Centering theory defines three data structures in order to describe the focus of a sentence: a list of forward-looking centers ( $C_f$ ), the preferred center ( $C_p$ ), and a single backward-looking center ( $C_b$ ).  $C_f$  indicates the salient items of the sentence that are focus candidates in the next sentence, and  $C_p$  indicates the most-preferred item of  $C_f$ .  $C_b$  describes the focus of a sentence with regard to the context of the previous sentences. Figure 45 provides a system overview that reveals how centers are selected.

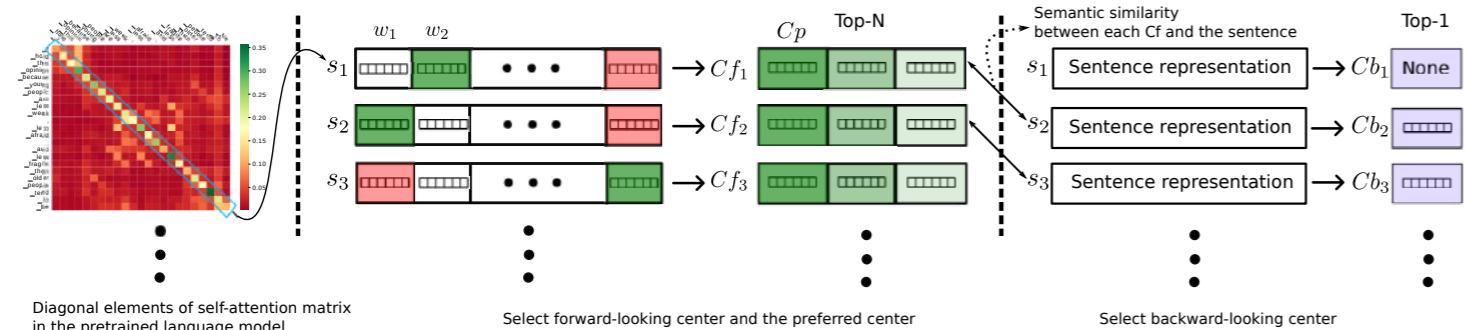


Figure 45: Selecting forward-looking centers ( $C_f$ ), preferred centers ( $C_p$ ), and backward-looking centers ( $C_b$ ).

To account for structural information, we propose using a structure-aware transformer that is built on top of a pretrained language model. This process creates trees that represent the structure of the text. We evaluate our model on two tasks: automated essay scoring and the assessment of writing quality. Results demonstrate that our model achieves state-of-the-art performance on both tasks. We examine the identified trees assigned to different essay scores and observe that texts of lower quality have deeper trees and more leaf nodes. These trees are also more skewed. Our findings suggest that the focus changes less frequently in texts of lower quality than in texts of higher quality.

Another interesting observation is that existing neural-network-based coherence-modeling systems process text differently from humans. These systems have access to all foci at the same time, and they track the changes to all foci simultaneously. The transformer – a recent AI technique – also relates all items simultaneously in order to capture semantic relations in a text. However, humans process text linearly and read sentences one-by-one and incrementally. Incremental processing has a direct connection to centering because processing a text sentence-by-sentence focuses the model's attention on a part of the

text, thereby making it easier to determine the focus of attention. To investigate this issue, we proposed a coherence model that interprets texts incrementally in order to capture lexical relations [Jeon and Strube, 2020b].

## NLP for database curation in the biomedical domain

### Mark-Christoph Müller

DeepCurate is a joint research effort of the HITS SDBV and NLP groups that was initiated in 2020. The goal of DeepCurate is to develop an integrated, machine-learning-based system for semi-automatic database

curation in the biomedical domain. The biocuration use case is provided by the SDBV group's SABIO-RK database and comprises (1) the selection of potential source-research papers for curation (paper

triage), (2) the identification of paper sections that contain relevant, curatable content, and (3) the extraction, normalization, and database insertion of this content.

**Natural Language Processing (NLP)** ist ein interdisziplinäres Forschungsgebiet, das mit Methoden der Informatik linguistische Fragestellungen bearbeitet. Die NLP Gruppe entwickelt Methoden, Algorithmen und Tools zur automatischen Analyse von Sprache. Sie konzentriert sich auf die Diskursverarbeitung und verwandte Anwendungen, wie zum Beispiel automatische Zusammenfassung und Lesbarkeitsbewertung.

2020 verteidigte Mohsen Mesgar seine Dissertation, die erste Verteidigung unter COVID-19-Bedingungen, bei der ein Teil des Komitees online teilnahm. Mohsen behandelte das Thema Textkohärenz in seiner Dissertation. Er wendete das von ihm entwickelte Modell auf die Bestimmung des Lesbarkeitsgrades von Texten an. Im Zuge der Entwicklung seines Modells wendete er zunächst traditionelle Methoden des maschinellen Lernens an, während er zum Ende seiner Arbeit mit guten Ergebnissen ein neuronales Netzwerk entwickelte, das gleichwohl auf linguistischen Einsichten beruhte. Mohsen arbeitet jetzt an der Technischen Universität Darmstadt als Postdoktorand. Der Masterstudent Lucas Rettenmeier reichte seine Masterarbeit an der Universität Heidelberg im Juli 2020 ein, interessanterweise an der Fakultät für Physik und Astronomie. Er untersuchte die Stabilität von Word Embeddings und wendete seine Ergebnisse auf das Bestimmen von linguistisch-semantischem Wandel an. Lucas arbeitet nach seinem Abschluss jetzt bei Amazon Web Services als Solutions Architect. Nebenbei lief er einen Halbmarathon beim (virtuellen) NCT-Run 2020 (siehe Kapitel 4) in weniger als einer Stunde und 20 Minuten!

Die NLP Gruppe publizierte weiterhin erfolgreich auf höchstem Niveau bei Konferenzen wie EMNLP und COLING, so etwa Sungho Jeon und Federico López. Federico absolvierte des weiteren ein dreimonatiges Praktikum bei Google Research in Mountain View, Kalifornien (von zu Hause).

2020 begannen wir zwei Kollaborationen mit HITS Gruppen. Zusammen mit der SDBV Gruppe starteten wir das vom BMBF geförderte Projekt DeepCurate mit dem Ziel das Kuratieren von Datenbankeinträgen in der biomedizinischen Domäne durch Computerhilfe zu unterstützen. Eine weitere faszinierende Kollaboration ergab sich aus den Arbeiten von Federico López zu hyperbolischen neuronalen Netzwerken für Entity Linking. Diese Arbeit stieß auf Interesse der GRG Gruppe, deren Arbeitsgebiet Differentialgeometrie ist. Diese Kollaboration führte 2020 schon zu einer Präsentation bei einem NeurIPS Workshop zu "Differential Geometry Meets Deep Learning" von Federico López, Beatrice Pozzetti, Steve Trettel (Stanford University) und Anna Wienhard. Wir werden weiterhin zusammen über Geometric Deep Learning arbeiten. Die NLP und die GRG Gruppen werden diese Zusammenarbeit im Rahmen eines HITS Lab-Projekts vertiefen (siehe Kapitel 2.6, S. 44/45).

Michael Strube war Program Co-Chair des "First Workshop on Computational Approaches to Discourse", der online im Rahmen der EMNLP 2020 stattfand. Der Workshop umfasste eingeladene Vorträge, Vorstellungen regulärer Forschungsbeiträge und eine große Podiumsdiskussion, deren Ziel das Etablieren einer Gemeinschaft von Wissenschaftlern ist, deren Forschungsthema Diskurs ist. Der zweite Workshop der Reihe ist schon in der Vorbereitung und wird im Rahmen der EMNLP 2021 stattfinden.

In the first phase of the project, the NLP group was active in several sub-tasks. To collect multi-modal paper-triage data (performed by the SDBV), two pools of candidate papers from two different biomedical domains had to be created.

Apart from the domain of general biochemical reactions and their kinetics, which constitutes the main domain of SABIO-RK, the second domain was related to the COVID-19 virus. The full text corpus for the first domain was downloaded from PubMed (<https://pubmed.ncbi.nlm.nih.gov/>), while the second was based on the CORD-19 dataset (<https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>). From these corpora, candidate papers for the manual triage process had to be selected automatically. In both cases, the selection was based on the presence of certain words and – in particular – of expressions of numerical measurements and their related units (mostly standard parameters from kinetics, such as  $K_{cat}$  and  $K_m$ ).

In order to support a more fine-grained string search involving regular expressions and proximity searches, the raw data were converted into the XML-based format of the annotation tool MMAX2 (<https://github.com/nlpATHits/MMAX2>). For both this conversion and the subsequent matching, an early version of the pyMMAX API [Müller, 2020] was

employed. As a result of this sub-task, the NLP group provided two collections of candidate papers to the SDBV, which were then used in the manual paper-triage-data collection.

A second, more-comprehensive effort involved the development of a component for a processing task called DB-to-document backprojection. In a nutshell, this task takes a single database record from SABIO-RK (or any other biomedical database) that represents, for example, a single measurement reported in a particular paper and attempts to determine the exact location in the paper from which the information was extracted during manual curation.

This task is challenging because the source papers are only available as digital images that were created by scanning the original paper print-outs. The decision to use these images as the basis for DB-to-document backprojection (as opposed to, e.g., PDF- or XML-based full texts, which are also available for many papers curated for SABIO-RK) stemmed from two main issues: First, the original paper printouts used for SABIO-RK curation contain manually added highlighting of important sections that we wanted to preserve, and second, future curation efforts (by SABIO-RK as well as by other biomedical databases) are likely to be paper-based

because printed paper is still the preferred medium for manual curation.

The ability to process scanned-paper images will thus continue to be important in the future. At the same time, an image (of sufficient quality) constitutes the most general and versatile representation of a scientific document that might also contain tables or non-textual information, such as graphs.

In the context of DeepCurate, DB-to-document backprojection was originally intended as a pre-processing step for creating training- and test data for machine-learning-based curation methods. While this remains the main focus of the project, DB-to-document backprojection is also useful as a tool for quality assurance. A first version of the DB-to-document-backprojection component is described in [Müller et al., 2020]. Figure 46 presents automatically created results in which database records (on the right) are linked to sections of the paper on the basis of matches in the OCR representation of the document image (read more about DeepCurate in Chapter 2.11, pp. 74/75).

troscopy. In all cases the expected carbon-13 upfield shift was observed for the deoxy carbon atom, and there was no evidence of any remaining starting material. The purity of the products was confirmed by comparison of the melting points, where crystalline, to reported values, and by thin-layer chromatography (EM silica plate, EtOH/water/ammonium hydroxide, 21/4/1). Detection was achieved by anisidine-HCl spray [26] with 10% sulfuric acid in methanol. All of the compounds yielded a single spot by TLC analysis, except for 6-deoxy-D-glucitol. The impurity that was detected in this compound was nonreactive to *p*-anisidine, and therefore was not a reducing sugar. This product mixture was purified on a AG-50WX8 equilibrated with 1 M  $\text{CaCl}_2$ . TLC analysis after chromatography revealed a single, non-reducing product whose identity was confirmed by NMR spectroscopy.

**Enzyme assays.**—Aldose reductase kinetics were examined in 100 mM phosphate buffer, pH 7.2, 27 °C, with 160  $\mu\text{M}$  NADPH ( $K_m = 1 \mu\text{M}$ ) by following the decrease in absorbance at 340 nm resulting from the oxidation of NADPH. Sorbitol dehydrogenase kinetics were determined in 100 mM HEPES buffer, pH 8, 30 °C, with 2 mM NAD ( $K_m = 11 \mu\text{M}$ ) by following the increase in absorbance at 340 nm resulting from the formation of NADH. One unit of activity is defined as 1  $\mu\text{mol}$  of product produced  $\text{min}^{-1} \text{mg}^{-1}$  of protein under the standard reaction conditions. The kinetic parameters for each substrate were determined by varying the substrate concentration at saturating concentration (> 20 times  $K_m$ ) of the cofactor and fitting the measured rates to the equation for Michaelis-Menten kinetics. Kinetic studies with a variety of substrates for each enzyme have shown that the affinity for the coenzyme is not significantly affected by the nature of the substrate [4,27]. When substrate inhibition was observed, the data were modeled by Eq. (1):

$$v = \frac{V_{max}[A]}{K_m + [A] + [A]^2/K_i} \quad (1)$$

where  $v$  is the measured rate, and  $K_m$  and  $K_i$  are the Michaelis and the substrate inhibition constants, respectively. The kinetic data were fit by using BASIC versions (Enzyme Kinetics

Package, SciTech International, Chicago, IL) of the kinetics programs originally written by Cleland [28].

**Spectroscopic studies.**—Circular dichroism studies were conducted on a JASCO J600 CD spectrometer. Samples were monitored near the maximum carbonyl ellipticity peak (285–290 nm) following a method for the quantitation of the proportion of acyclic form [29]. Temperature studies were performed with a thermally heated cuvette block. Spectra were recorded when samples were equilibrated at each temperature, as indicated by no further changes in the measured ellipticities. The ellipticities (in mdeg) were reproducible, leading to less than a 10% error in the calculated percent acyclic form.

### 3. Results

**Substrate specificity of aldose reductase.**—ALR catalyzes the reduction of the physiological substrate, D-glucose, to D-glucitol. Structural analogs of D-glucose, in which one of the sugar hydroxyl groups was replaced with either a fluorine or a hydrogen, were examined as possible alternative substrates for the reaction catalyzed by ALR. ALR binds only the acyclic form of glucose and does not catalyze ring-opening [30]. Therefore it is necessary to correct the substrate concentrations for the amount of acyclic form present in order to compare the relative kinetic parameters of these alternative substrates. Table 1 gives the percent acyclic form of these substrates, as determined by circular dichroism measurements. The 4-fluoro analog contains 2.5 times the amount of acyclic form, and the 3-fluoro analog about the same percentage as glucose. The 6-deoxy-D-glucose has been determined to have 0.002% acyclic form [29], while 2-deoxy-D-glucose has about 2.5 times this amount.

Given the broad specificity of ALR, it was not surprising that all of the fluoro- and deoxysugar analogs that were tested were found to be substrates for the enzyme. However, unexpectedly, 3-fluoro- and 4-fluoro-D-glucose were found to be better substrates, with significantly lower  $K_m$  values and higher  $k_{cat}/K_m$

**ID: 159982103 Max. score 1.000 @ sec size 6**  
100 mM phosphate buffer, pH 7.2,  
['pH'] --> pH  
['7.2', '7.2'] --> 7.2  
['100', '100.0', '100.0'] --> 100  
['mM'] --> mM  
['phosphate'] --> phosphate

**ID: 159982102 Max. score 1.000 @ sec size 9**  
100 mM phosphate buffer, pH 7.2, 27 °C, with  
['27', '27.0', '27.0'] --> 27  
['°C'] --> °C  
['100', '100.0', '100.0'] --> 100  
['mM'] --> mM  
['phosphate'] --> phosphate

**ID: 101200 Max. score 1.000 @ sec size 6**  
NAD ( $K_m = 11 \mu\text{M}$ ) by  
['μM', 'pM'] --> pM  
['Km', 'K<sub>m</sub>', 'K<sub>m</sub>', 'K<sub>m</sub>', 'K<sub>m</sub>', 'K<sub>m</sub>', 'K<sub>m</sub>', 'K<sub>m</sub>', 'Ktn', 'Kni'] --> Km  
['11', '11.0', '11.0'] --> 11  
['NAD+', 'NAD', 'DPN', 'Nadide', 'beta-NAD+'] --> NAD

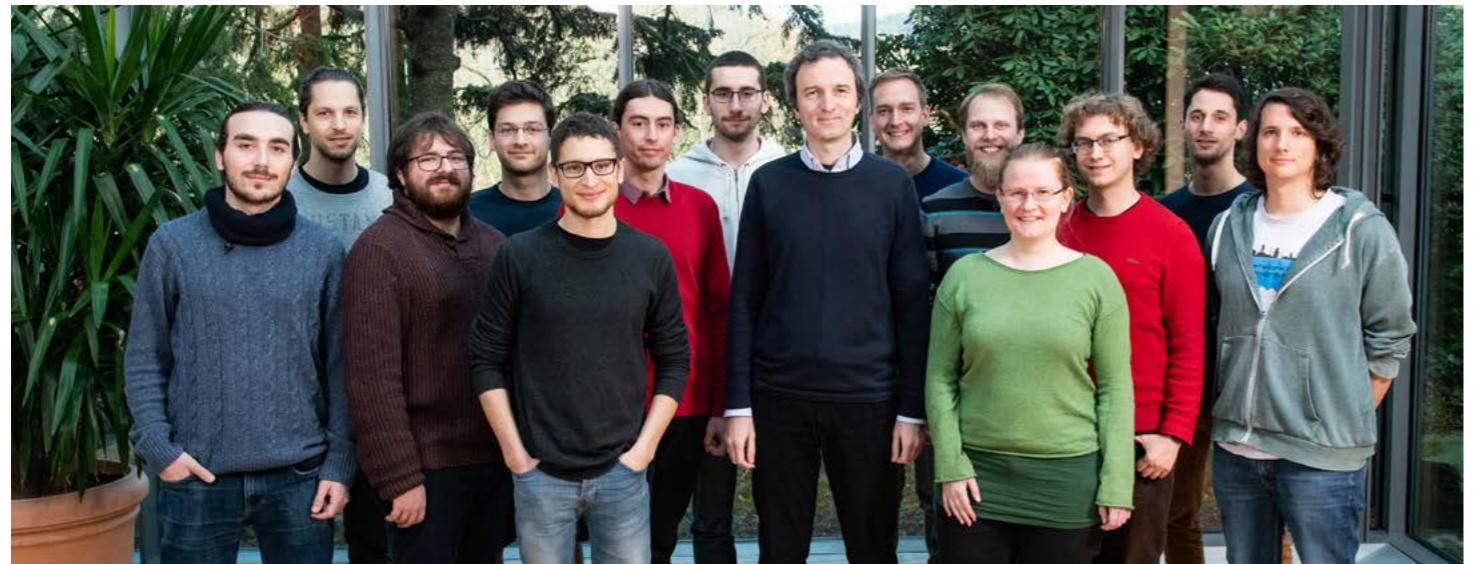
**ID: 101200 Max. score 1.000 @ sec size 6**  
mm NAD ( $K_m = 11 \mu\text{M}$ )  
['μM', 'pM'] --> pM  
['Km', 'K<sub>m</sub>', 'K<sub>m</sub>', 'K<sub>m</sub>', 'K<sub>m</sub>', 'K<sub>m</sub>', 'K<sub>m</sub>', 'K<sub>m</sub>', 'Ktn', 'Kni'] --> Km  
['11', '11.0', '11.0'] --> 11  
['NAD+', 'NAD', 'DPN', 'Nadide', 'beta-NAD+'] --> NAD

**ID: 101105 Max. score 1.000 @ sec size 3**  
160 pM NADPH  
['μM', 'pM'] --> pM  
['160', '160.0', '160.0'] --> 160  
['NADPH', 'TPNH', 'beta-NADPH'] --> NADPH

Figure 46: OCR representation of a scanned document (left) and automatically extracted database records (right).

# 2 Research

## 2.10 Physics of Stellar Objects (PSO)



### Group Leader

Prof. Dr. Friedrich Röpke

### Staff members

Dr. Róbert Andrásy  
Dr. Johann Higl  
Javier Morán Fraile

### Scholarship holders

Leonhard Horst (HITS Scholarship)  
Christian Sand (HITS Scholarship)

### Visiting scientists

Sabrina Gronow  
Florian Lach  
Giovanni Leidi  
Dr. Fabian Schneider (Gliese Fellowship)  
Theodoros Soulantasis (IMPRS PhD Student at MPIA Heidelberg)

### Students

Dave Bubeck  
Manuel Kramer  
Melvin Moreno  
Patrick Ondratschek

"We are stardust" – the matter we are made of is largely the result of processing the primordial material formed during the Big Bang. All heavier elements originate from nucleosynthesis in stars and in gigantic stellar explosions. How this material formed and how it is distributed throughout the Universe are fundamental concerns 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 and are the fundamental building blocks of galaxies and larger cosmological structures. With the help of extensive numerical simulations, the

Physics of Stellar Objects research group seeks to understand the processes that take place in stars and stellar explosions. Newly developed numerical techniques and the ever-increasing power of supercomputers facilitate the modeling of stellar objects in unprecedented detail and with unparalleled precision. One of our group's primary goals is to model the thermo-nuclear explosions of white dwarf stars that lead to the astronomical phenomenon known as Type Ia supernovae. These supernovae are the main source of iron in the Universe and have been instrumental as distance indicators in cosmology, which has led to the spectacular

discovery of the accelerating expansion of the Universe. Multi-dimensional fluid-dynamic simulations in combination with nucleosynthesis calculations and radiative-transfer modeling provide a detailed picture of the physical processes that take place in Type Ia supernovae and are also applied in the PSO group to other kinds of cosmic explosions. Classical astrophysical theory describes stars as one-dimensional objects in hydrostatic equilibrium, an approach that has proven extremely successful and

explains why stars are observed in different configurations while also providing a qualitative understanding of stellar evolution. However, simplifying assumptions limit the predictive power of such models. Using newly developed numerical tools, our group explores dynamic phases in stellar evolution via three-dimensional simulations. Our aim is to construct a new generation of stellar models based on an improved description of the physical processes that take place in stars.

### Shaking up stars

Internal waves are hydrodynamic perturbations that propagate inside a stratified medium. They are a well-known phenomenon in the fields of oceanography and atmospheric science and are responsible for local weather effects (e.g., föhn winds) and global circulations in the atmosphere (e.g., quasi-biennial oscillation). One mechanism by which such waves are generated is via convective motions in the upper atmosphere or in the upper layers of oceans. Convection is usually restricted to some fixed part of the atmosphere. When convective plumes reach the boundary of a convective region, they "hammer" against it and excite waves in the adjacent, stable layer. The same mechanism is also active in stars. An example involves stars with masses of more than twice that of the Sun and that have a convective core and a stable envelope. In these stars, waves are excited at the core boundary and are expected to propagate all the way to the surface of the star, where they can be detected as variations in brightness.

### Asteroseismology

The field of asteroseismology uses these variations in brightness to infer the interior structure of stars in order to improve one-dimensional stellar models. In general, two types of waves can be observed: gravity waves, for which buoyancy acts as the restoring force and that have a rather low oscillation frequency, and

sound waves, which are triggered by pressure fluctuations and can be observed as higher-frequency signals. In a one-dimensional model, the excitation of waves via convection needs to be approximated since convection is an intrinsically multi-dimensional process. Moreover, one-dimensional models cannot predict the amplitude of waves. However, depending on the amplitude, non-linear effects can lead to a redistribution of angular momentum and to a mixing of chemical elements throughout a star. Both effects can have a significant impact on the further evolution of the star. Multi-dimensional numerical simulations are therefore needed to verify the currently used models of wave excitation and to estimate wave amplitudes.

### Flows in deep stellar interiors

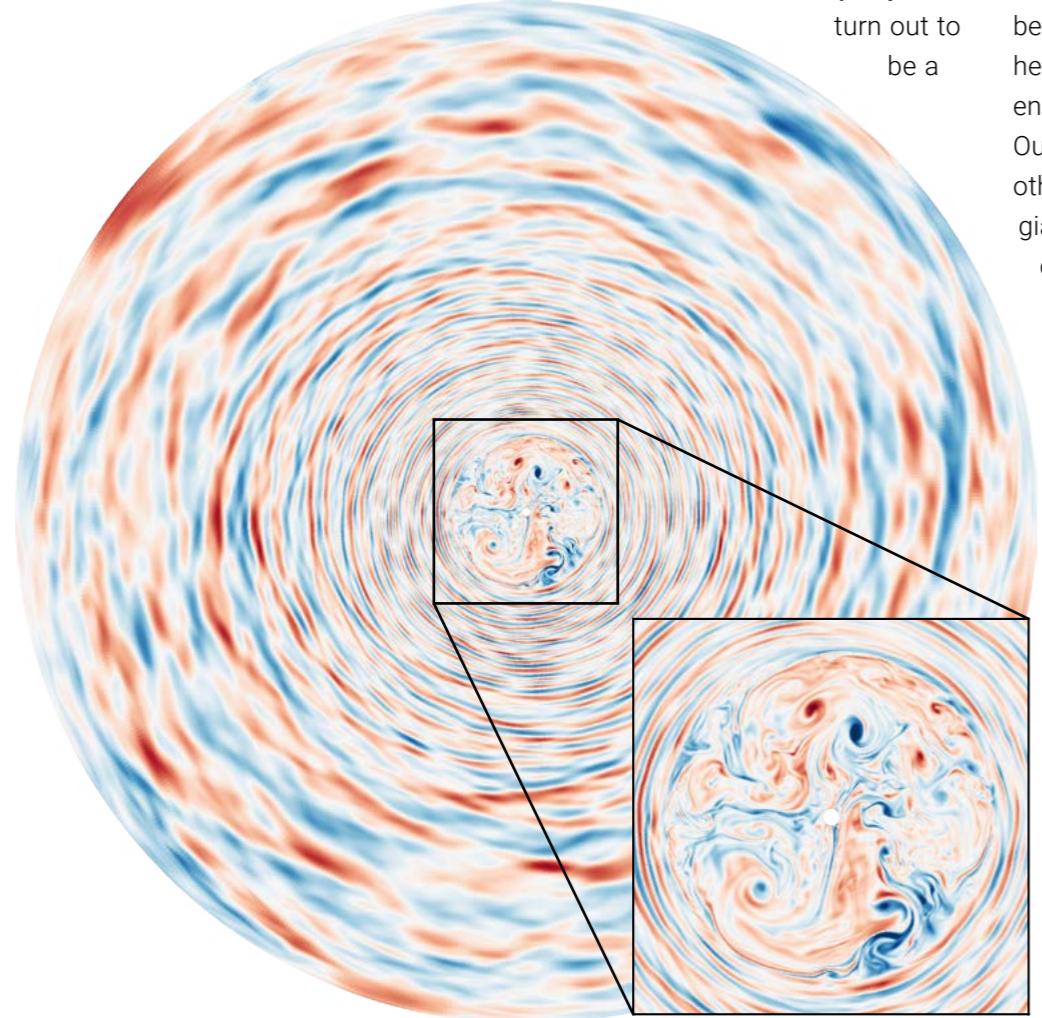
Flows in the interior of stars usually have a very low Mach number – that is, the flow speed is usually several orders of magnitude less than the local speed of sound. The frequency of gravity waves is related to the flow speed. In order to follow the propagation of gravity waves, it is therefore sufficient to follow the flow. Pressure waves, in contrast, require that a simulation also be accurate on the much smaller timescales related to the speed of sound. This large separation of timescales makes it highly challenging to simulate both gravity- and pressure waves in the same simulation.

The SLH code, developed in the PSO group at HITS, was designed to deal with this special situation based on a set of numerical techniques that remove numerical artifacts that are common in simulations of low-Mach-number flows. At the same time, the SLH code avoids approximations of the hydrodynamical equations that are often employed in simulations of low-Mach-number flows but prevent the resolution of sound waves.

### Simulations of wave generation

The capabilities of SLH have been thoroughly tested by the PSO group [Horst et al., 2020] and demonstrate the superiority of SLH compared with standard numerical schemes in following the propagation of gravity waves at very low Mach numbers. The group's publication also revealed that SLH is indeed capable of simultaneously simulating the excitation and propagation of gravity- and pressure waves, as was demonstrated with the example of a realistic stellar model of a 3-solar-mass star for which the evolution of the convective core and a large fraction of the stable envelope were followed for about 1 month of physical time in a two-dimensional simulation. The convective motion excites a rich spectrum of waves in the stable layer. A careful analysis of the waves revealed that their properties were in excellent agreement with theoretical predictions. In contrast to one-dimensional models, wave amplitudes can be easily extracted from the simulations. While the results hint at

possible non-linear wave behavior, it is not possible to make strong predictions with respect to mixing or to angular-momentum transport due to the reduced dimensionality of the simulations (see Fig. 47). Future three-dimensional simulations should be able to improve on this point and are currently being prepared in the PSO group.



*Figure 47: Snapshot of the internal-wave pattern in a two-dimensional simulation of a 3-solar-mass star. The color coding shows temperature fluctuations, with red and blue indicating temperatures below and above the angularly averaged temperature, respectively. The inset magnifies the core region, where the flow is convective (see [Horst et al., 2020]).*

## Stellar Dance

### Lonely stars?

Although on clear nights the sky appears crowded, the stars we observe are often extremely far apart from one another. The resolving power of modern telescopes, however, changes this picture. In fact, many stars that appear to be solitary objects turn out to be a

requires fuel, and after some time, this fuel is exhausted. At this point, the stellar structure rearranges such that new nuclear reactions commence and equilibrium is recovered. Our Sun will thus evolve through giant stages. As a so-called asymptotic-giant-branch (AGB) star, it will develop a pronounced a core-envelope structure in which a central and largely inert carbon/oxygen core will be embedded in shells with burning helium/hydrogen and a very large envelope.

Our Sun is a lonely star, but many others are not. Before reaching the giant stages, stars may have close companions.

### Common-envelope evolution

One of the unsolved problems of stellar astrophysics is the question of what happens to the system when the faster-evolving, more-massive ("primary") star evolves into a giant. It is bound to catch its companion in a "common envelope," inside of which the companion and the core of the primary star orbit each other. Because this happens in the envelope gas, tidal friction drains kinetic energy from the orbital motion and transfers it to the energy of the envelope gas,

which has two fundamental consequences: The separation of the stellar cores shrinks, and the envelope material may be ejected from the system, eventually forming a so-called planetary nebula. The leftover tight binary consisting of two stellar cores may engage in interactions that – depending on the nature of the two initial stars – can give rise to spectacular astrophysical phenomena, such as supernova explosions.

pair of companions. Stars are also not as changeless as it first may seem. Indeed, they have been found to evolve through distinct phases that are characterized by vastly different sizes and internal structures. The reason for this evolution lies in the interplay between gravity, which pulls the stellar material together, and nuclear reactions in the stars' interiors, which produce energy such that thermal pressure can balance gravitational pull. Nuclear burning, however,

"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. Wie dieses Material gebildet wurde und wie es sich im Universum verteilt, stellen für Astrophysiker fundamentale Fragen dar.

Sterne sind fundamentale Bausteine von Galaxien und aller größeren kosmologischen Strukturen. Gleichzeitig machen stellare Objekte das Universum für uns in astronomischen Beobachtungen überhaupt erst sichtbar. Sterne scheinen im optischen und anderen Teilen des elektromagnetischen Spektrums. Am Ende ihrer Entwicklung kollabieren massereiche Sterne zu Neutronensternen oder Schwarzen Löcher. Eine Verschmelzung solcher kompakten Objekte wurde kürzlich mit Hilfe von Gravitationswellen beobachtet, die ein neues Fenster für astronomische Beobachtungen des Universums öffnen.

Unsere Forschungsgruppe **Physik stellarer Objekte (PSO)** strebt mit Hilfe von aufwendigen numerischen Simulationen ein Verständnis der Prozesse in Sternen undstellaren 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. Die klassische astrophysikalische Theorie beschreibt Sterne als eindimensionale Objekte im hydrostatischen Gleichgewicht. Dieser Ansatz ist extrem erfolgreich. Er erklärt, warum wir Sterne in verschiedenen Konfigurationen beobachten, und liefert ein qualitatives Verständnis der Sternentwicklung. Die hierbei verwendeten vereinfachenden Annahmen schränken jedoch die Vorhersagekraft solcher Modelle stark ein. Mit neu entwickelten

numerischen Hilfsmitteln untersucht unsere Gruppe dynamische Phasen der Sternentwicklung in dreidimensionalen Simulationen. Unser Ziel ist es, eine neue Generation von Sternmodellen zu schaffen, die auf einer verbesserten Beschreibung der in ihnen ablaufenden physikalischen Prozesse basiert.

Eine weitere Komplikation, die in klassischen Sternentwicklungsmodellen nur sehr grob angenähert werden kann, ist die Binarietà. Wohl wegen des Beispiels unserer Sonne tendieren wir oft dazu, Sterne als isolierte Objekte zu sehen; tatsächlich findet man die meisten von ihnen jedoch in Systemen mit zwei oder sogar mehr Sternen. Einige von diesen wechselwirken miteinander, und das hat weitreichende Auswirkungen auf ihre weitere Entwicklung. Solche Interaktionen sind inhärent mehrdimensional und können in klassischen Modellen nicht konsistent behandelt werden. Die PSO-Gruppe führt dreidimensionale Simulationen zu stellaren Wechselwirkungen durch, um neue Einsichten in diese entscheidenden Phasen der Entwicklung von Sternensystemen zu gewinnen.

Das dritte Forschungsfeld der PSO Gruppe ist die Modellierung von thermonuklearen Explosionen Weißer Zwerge, 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 strömungsdynamische 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 Explosio-nen angewendet.

### 3D hydrodynamical simulations

Common-envelope interaction is an inherently dynamic and multidimensional phase in binary stellar evolution – that is, it cannot be described in classical one-dimensional stellar-evolution models. In order to study the physical process in detail, the PSO group developed a modeling approach based on the AREPO code originally developed by Volker Springel and his team at HITS (see former Annual Reports). This code is particularly well-suited to modeling common-envelope interaction because it

follows fluid dynamics on a moving computational mesh, which has many advantages over conventional static grids for the problem at hand.

### Common envelopes resulting from AGB primary stars

Systems that enter common-envelope evolution when the primary star becomes an AGB star are of particular interest. Once the envelope is ejected, its core – a compact white dwarf consisting of carbon and oxygen – is left with a close companion. Such systems form many of the

observed planetary nebulae and are thought to be the progenitors of Type Ia supernovae. Compared with other giant stages, however, the AGB phase is numerically challenging. The envelope is extremely bloated and only slightly bound, which requires meticulous care in the setup to avoid non-physical perturbations of the delicate equilibrium. Common-envelope interactions with AGB primary stars have therefore not often been simulated in the past. With its new tools, the PSO group tackled this problem and was able to perform

corresponding simulations [Sand et al., 2020]. The snapshots shown in Figure 48 were taken from a system with a 6-billion-year-old AGB primary star that – at the onset of common-envelope interaction with a companion – had a radius 200 times as large as that of the Sun. Its core is marked with an "x." The companion star, marked with a "+," has the mass of our Sun. It was initialized in an orbit that is not in exact co-rotation with the envelope, and its inspiral was caused by drag force.

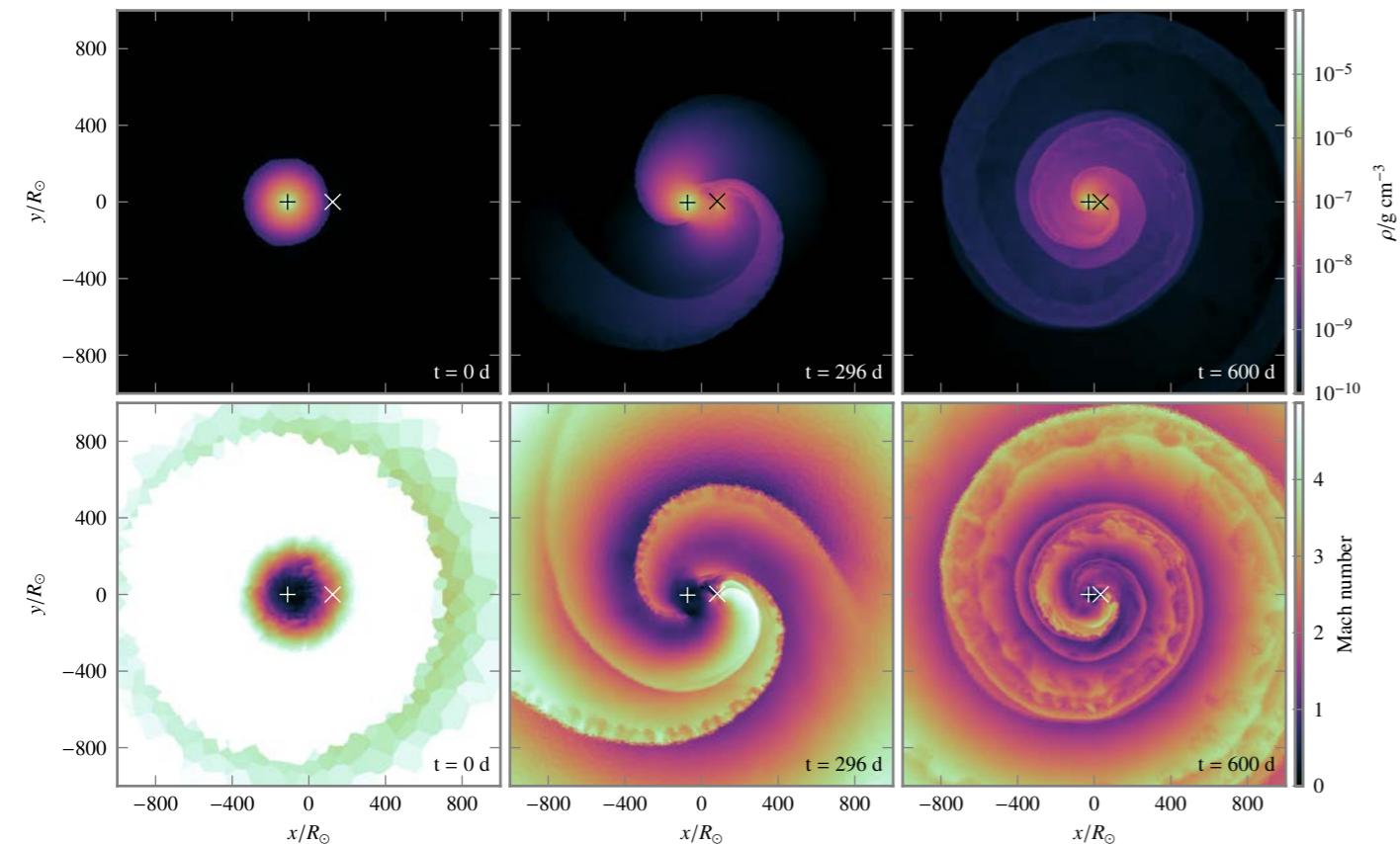


Figure 48: Common-envelope evolution with an AGB primary star: gas density (upper row) and Mach number (the ratio of fluid velocity to sound speed; lower row) in the orbital plane are plotted at the beginning of the simulation as well as after both one and four orbits. The initial centers ("cores") of the giant star and the companion are marked by a plus symbol and a cross symbol, respectively. The first spiral arm forms where the companion dives into the envelope, and another is formed on the opposite side. Large-scale instabilities between layers can be identified particularly well in the lower Mach plot [Sand et al., 2020].

While the first orbit takes about 300 days, another three orbits occur within the next 300 days as the companion spirals in and approaches the core of the AGB star. The entire event takes only a few years before

the system settles on a final orbit. An important (and challenging) question is whether the envelope material is ejected in this common-envelope phase. Where does the energy for the ejection come from? The orbital energy of the binary is released during the inspiral. Although the envelope of the AGB star is quite loosely bound, the energy from the inspiral of the companion – which is eventually transferred to the envelope material – is insufficient to eject and remove it completely. A second and

This energy acts in spiral arms, boosting the expansion (see Figure 49).

When including this mechanism, the simulations performed by the PSO group indicate full envelope ejection (see Figure 50).

By varying the parameters of the initial system, such simulations have also revealed that the efficiency of mass loss and the final orbital separation of the core binary system

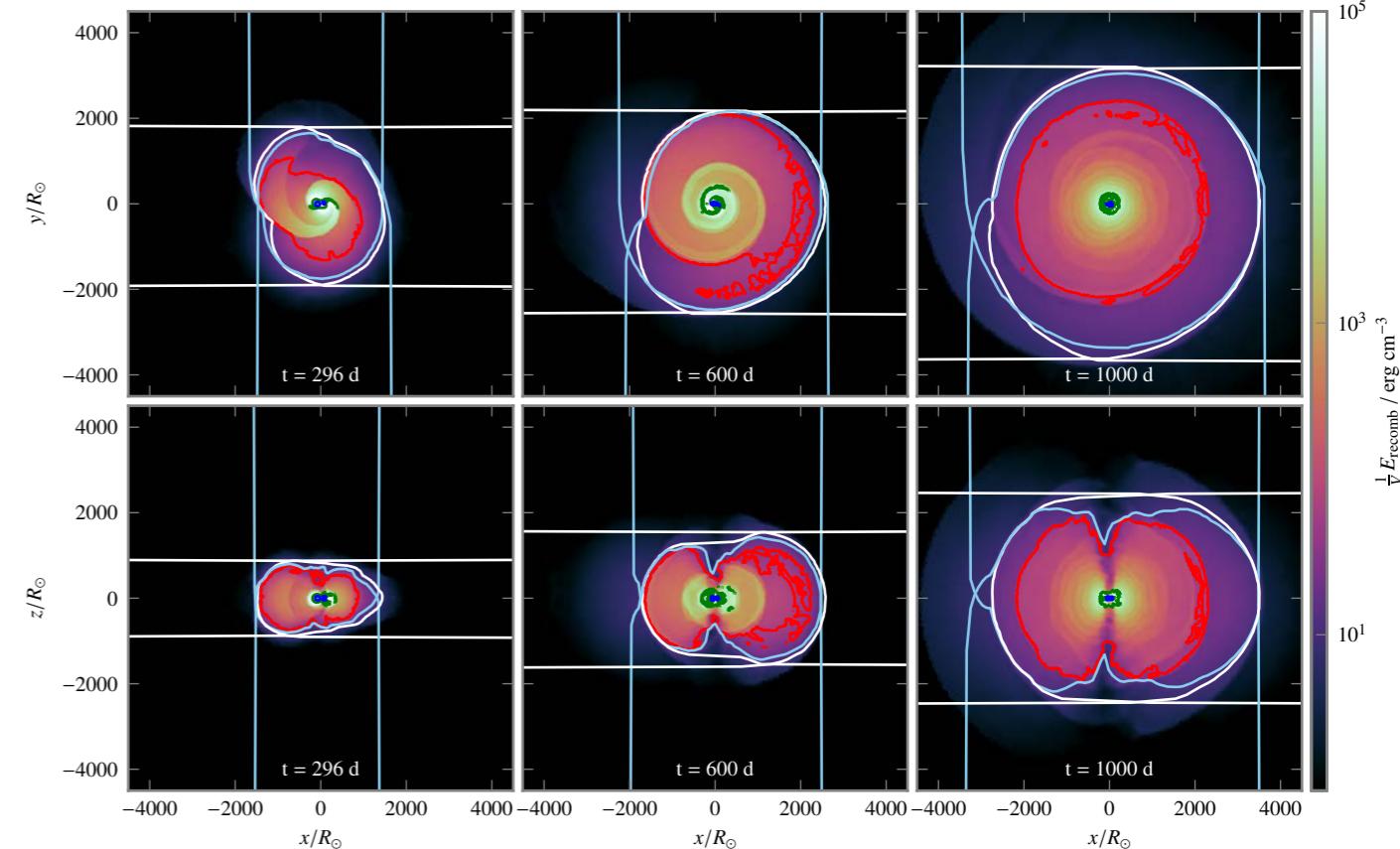


Figure 49: The ionization energy density inside the photosphere (white and light-blue lines) can be used to remove the envelope. The inner contours enclose regions with high ionization fractions [Sand et al., 2020].

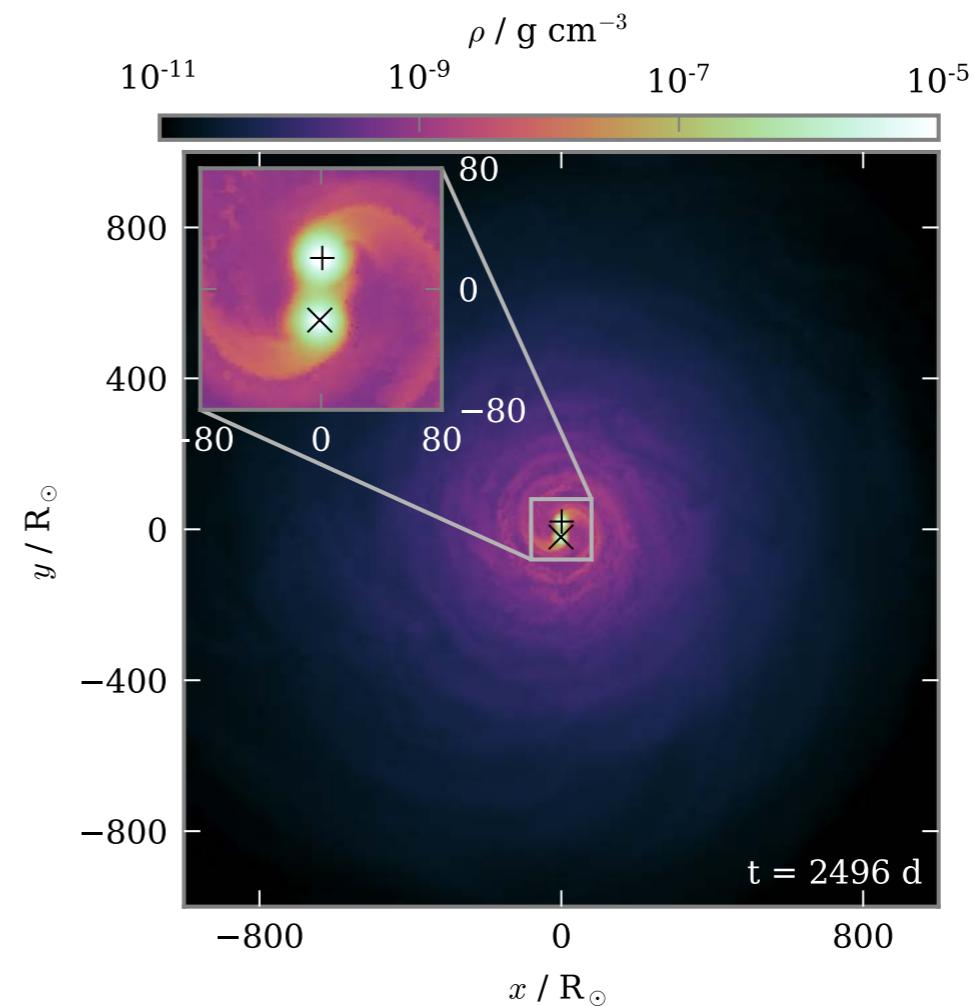


Figure 50: Gas density in the orbital plane at the end of our simulation (after about seven years). The cores of the two stars (marked by the symbols) perform a close dance while preserving a spiral structure around them [Fig. 4 of [Sand et al., 2020]].

important effect that powers envelope ejection is the release of the ionization energy of the material in recombination processes: When gas expands and cools, ions recombine to form atoms and release energy.

depend on the mass ratio of the two stars. This finding is currently being further explored by the PSO group with simulations of common-envelope phases in systems involving stars of different natures.

# 2 Research

## 2.11 Scientific Databases and Visualization (SDBV)



### Group Leader

PD Dr. Wolfgang Müller

### Staff members

Helen Desmond (since July 2020)

Dr. Dorotea Dudaš

Dr. Sucheta Ghosh

Martin Golebiewski

Xiaoming Hu

Vivien Louise Junker (since October 2020)

Dr. Olga Krebs

Michael Lieser (since October 2020)

Ghadeer Mobasher (since January 2020)

Ina Pöhner

Dr. Maja Rey

Dr. Natalia Simous

Dr. Andreas Weidemann

Benjamin Winter

Dr. Ulrike Wittig

This year we highlight our COVID-19 activities, DeepCurate, our HITS Lab project, and NFDI4Health, the German National Research Data Infrastructure for Personal Health Data. We begin, however, by summarizing the group's focus and its main development.

The SDBV group aims to provide services involving FAIR data. The FAIR acronym stands for Findable, Accessible, Interoperable, and Reusable data. Using the SEEK system, we help people make their own data FAIR. In SABIO-RK, our professional curators provide FAIR data on enzyme-catalyzed reactions extracted from the literature. FAIR data are computationally discoverable. Ideally, a smart harvester should recognize FAIR data as such and be able to harvest them. Making these ideas work implies standardization. Participating in standardization efforts – such as COMBINE, STRENDA, and ISO/TC 276 Biotechnology – constitutes a significant part of our work. All of our work benefits from the close collaboration of biologists and biochemists with computer scientists.

Most of our work is project-funded. We have three types of projects: (i) projects with a biological/biomedical focus and for which the SDBV group is an infrastructure partner, such as LiSyM; (ii) collaborative infrastructure projects with the SDBV group, such as FAIRDOM and de.NBI; and (iii) research projects, such as DeepCurate and PoLiMeR.

In the past year, one project that (along with its forerunners) had helped to shape the group's previous decade was completed: FAIRDOM, a project focused on FAIR Data, Operations, and Models. FAIRDOM and its derivatives have been featured in the SDBV group in recent years, and FAIRDOM work constitutes part of two of the new projects that began last year: ELIXIR CONVERGE and NFDI4Health.

We hope that our contribution to the NFDI4Health consortium (which began in 2020) and to NFDI4Biodiversity will further develop the work of FAIRDOM.

### COVID-19 activities

With the services developed in our group, namely SABIO-RK and FAIRDOM-SEEK, we were able to react quickly to the new scientific challenges brought about by the SARS-CoV-2 coronavirus pandemic and the new COVID-19 disease. FAIRDOM-SEEK is used as a flexible platform for storing and sharing

research data within national and international COVID-19 research communities.

### COVID-19 Disease Map in the FAIRDOMHub

The international COVID-19 Disease Map initiative is an effort involving more than 230 scientists with the goal of building a comprehensive, standardized knowledge repository

of SARS-CoV-2 virus–host interaction mechanisms guided by input from domain experts and based on published work. The resulting map is a platform for visual exploration and computational analyses of the molecular processes involved in SARS-CoV-2 entry, replication, and host–pathogen interactions as well as immune response and host-cell recovery- and repair mechanisms.

Figure 51: Screenshot of the COVID-19 Disease Map project space (upper panel) on the SEEK-based FAIRDOMHub (<https://fairdom-hub.org/projects/190>) with two examples of SARS-CoV-2 models registered in FAIRDOMHub (lower panels).

The map supports the research community and improves our understanding of the disease with the aim of facilitating the development of efficient diagnostics and therapies. We support the COVID-19 Disease Map community with a project space in our public research-data-management platform, FAIRDOMHub (<https://fairdomhub.org/projects/190>), which groups project contributors, data, computational models, and literature as well as data and curation guidelines for the community (Fig. 51, previous page). FAIRDOMHub and its underlying SEEK software have been under development by the SDBV group for over 10 years.

#### **COVID-19 Study Hub based on SEEK for clinical and epidemiological studies**

Due to the pressing need during the progressing pandemic, a quickly growing number of clinical and epidemiological COVID-19 studies are planned or already ongoing, but there is a lack of coordination among these efforts for securing common standards, comparable results, and – most importantly – unified access to these results. Several partners in the German National Research Data Infrastructure for Personal Health Data (NFDI-4Health), including the SDBV group, are developing a COVID-19 Study Hub for Germany with funding from the DFG as a COVID-19 task force (<https://www.nfdi4health.de/index.php/task-force-covid-19-2/>). This publicly accessible platform bundles information on relevant clinical and epidemiological studies in Germany, their publicly available study documents and results (e.g., measured parameters), and other information about the studies.

The COVID-19 Study Hub is partially

based on the SEEK software in an effort to support data storage and exchange. SEEK is used to store study documents and make them accessible. These documents include study-protocol templates and data dictionaries as well as information on study-metadata structures – such as data models that describe study subjects and their clinical parameters – in addition to treatment outcomes and similar information (<https://seek.studyhub.nfdi4health.de>). Additionally, direct links to primary resources and websites for the studies are included. Using the SEEK web services, this information also directly feeds a COVID-19 study search portal implemented by our project partners at ZB MED in Cologne (<https://covid19.studyhub.nfdi4health.de>). The standardization of metadata that describe the studies as well as their subjects and results also constitute part of our activities in the NFDI4Health COVID-19 task force and contribute to the aim of making the studies and their content "FAIR."

#### **COVID-19 kinetics data extraction for SABIO-RK**

Within our SABIO-RK database (<http://sabiork.h-its.org/>), users can find biochemical reactions together with the catalyzing enzymes and kinetic constants that describe reaction velocities ( $V_{max}$ ,  $k_{cat}$ ) as well as concentrations of reaction participants for half-maximal activities ( $K_m$ ) or of inhibitors for half-maximal inhibitions ( $K_i$ ,  $I_C50$ ). Such structured, annotated, and standardized collections of kinetic constants that originate from scientific publications are needed not only to model metabolic networks but also for basic research, for example, in drug development.

Therefore, in 2020, we put effort into manually extracting from the literature such inhibition constants and other kinetic parameters involved in SARS-CoV-2 and related viruses – such as MERS and SARS-CoV – with the aim of supporting the search for treatments against COVID-19. Many of these potentially antiviral agents are directed (1) against the spike protein needed by the virus to enter the host cell or (2) against the viral proteases (3CLpro, PLpro) needed for the virus's life cycle.

#### **DeepCurate**

DeepCurate represents a step toward semi-automated data curation and offers a deeper understanding of the biocuration process. Within the DeepCurate project, which has been funded by the BMBF since January 2020, the database aims to strengthen biological-data curation using semi-automatic processes within the biocuration framework of SABIO-RK via collaboration with the NLP group at HITS. The SABIO-RK database for biochemical reaction kinetics has successfully served the community for several years. DeepCurate also seeks to provide a deeper understanding of the curators' cognitive and ergonomic processes, which can reveal their perceived workload and lead to a better understanding of points of difficulties or subtleties of comprehension. We are currently looking deeply into the document-triage procedure. Document triage is defined as the process of pooling candidate papers from a large journal repository – such as PubMed – for later curation in the SABIO-RK database. This pooling is performed by the curators via

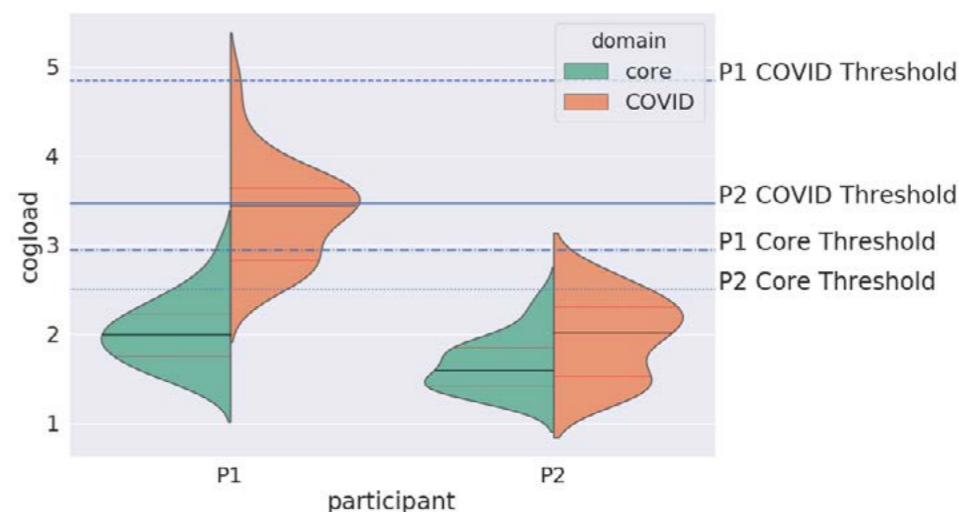


Figure 52: Threshold computation for cognitive load (PD in %) using quartile estimation (P1: Participant 1, P2: Participant 2).

manual searches using a domain-specific ontology and a parameter list from working memory. The curators accustom themselves to the domain-specific ontology from the core or from the habitual domain of their work.

Another focus of the project is the analysis of the effects of working with documents from the new COVID-19 domain. We found that it takes a statistically significant and greater amount of effort to work with documents from the new COVID-19 domain than with those from the core domain. We used the cognitive load (that is, pupil dilation [PD] in percentage) to measure the effort and then proposed a threshold for effortless reading using quartile-based outlier estimation of the cognitive load (Fig. 52).

We can thus now send feedback to the curators in the case of a difficult curation task, and we can also determine the curation cost using this metric. We further observed that the curators took around twice as long to complete the triage of COVID-domain literature than of core-domain literature. We investigated whether the curators' familiarity with the curation in the new domain increased with the number

of completed tasks. In this query, the term familiarity implied attribution to long-term memory storage and included both accommodation with the screen-based triage setup using an eye tracker and the ease of reading papers from a new domain.

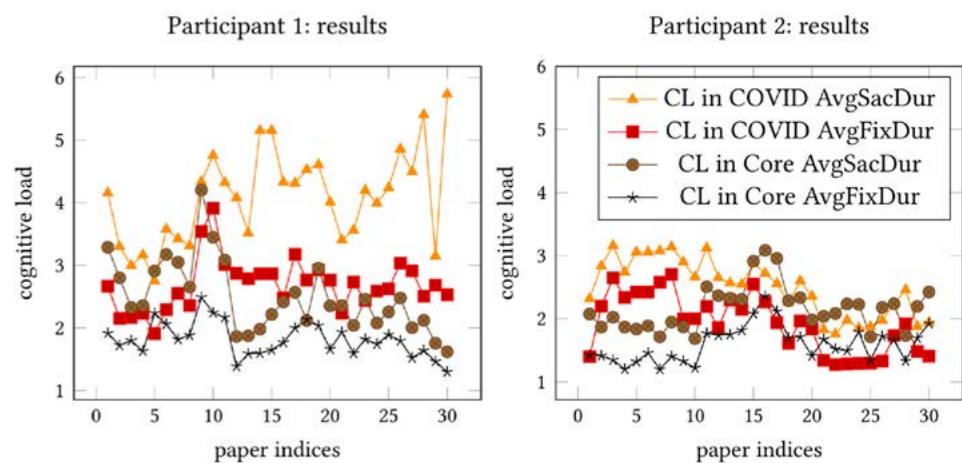


Figure 53: Cognitive load (PD in %) for fixation and saccade for two domains for both participants (same legend for both figures).

We noted that the cognitive load related to eye fixations decreased and remained stable over time for the new domain, which could represent weak evidence of increasing familiarity with the new domain. The fixations were eye movements in which the retina stabilized over a stationary object of interest used in reading. The rapid motion of the eye

from one fixation to another is called a saccade and is useful in visual search. In Figure 53, we observed that after working with the 15th of 30 papers, the cognitive load during the fixations decreased, whereas the cognitive load for the saccade also decreased for one of the participants after the 15th task. We also cross-checked the curators' responses and found that this decrease was not related to the difficulty of the respective paper-triage task. This study revealed that the triage procedure includes both reading and a visual search, which are comparable in terms of time. In our case, the visual search required greater effort (that is, a greater cognitive load) than did the reading. It is also possible that the curators needed to search less and to read

comparatively more for the familiar domain because the structure of the paper may have been familiar to them and they may have understood where the required data could be found. We look forward to using the proposed metrics in the main curation task in the future (read more about DeepCurate in Chapter 2.9, pp. 63-65).

## German National Research Data Infrastructure for Personal Health Data (NFDI4Health)

The largest change to the SDBV group in 2020 was the launch of NFDI4Health – the German National Research Data Infrastructure for Personal Health Data – in October 2020 (<https://www.nfdi4health.de>). As one of 18 partner institutions in the initiative, the SDBV group at HITS brings its many years of expertise with scientific databases and data standardization to the table. Within the framework of the German National Research Infrastructure (NFDI), the multidisciplinary team of NFDI4Health is set to establish a research-data infrastructure for personal-health data in Germany. The project – along with 8 other German research-data infrastructures – is funded by both the German Federal Government and German state governments. The SDBV group is directly involved in the initiative as a key partner and assumes a leading role in the data standardization. The SDBV group will also make its SEEK software suite available to the initiative as part of the infrastructure architecture that will be built by NFDI4Health over the next few years.

NFDI4Health provides a long-term perspective for the SDBV group as this infrastructure is intended to be sustainably funded, with an initial runtime of 5 years. NFDI4Health represents an interdisciplinary research community that integrates major German institutions with experience as health-data collectors and holders, health-data managers and analysts, as well as methodology developers. In addition to the 18 partners funded by the initiative, a total of 46 renowned institutions

from the health sector have confirmed their participation, including major professional associations and epidemiological cohorts. Letters of support have been received from 37 national and international institutions.

The project will pursue the following main objectives: (1) to implement a federated health-data infrastructure in Germany for searching for and accessing healthcare data and health databases, (2) to enhance data sharing and the data linkage of personal-health data in compliance with privacy regulations and ethics principles, (3) to enable the development and deployment of new consent-management mechanisms and augmented data-access services, and (4) to foster data sharing and cooperation between the communities involved in clinical research, epidemiological health, and public health.

Cohort studies, health-surveillance systems, and clinical trials are characterized by a deep phenotyping of study subjects (healthy individuals and/or patients) via questionnaires, medical examinations, molecular and genetic profiling, and collecting relevant metadata on study subjects. Their longitudinal nature and high quality make these data and their metadata a valuable research resource for developing preventive and therapeutic measures at the individual- and population levels. Rendering these types of health data FAIR (Findable, Accessible, Interoperable, and Reusable) by providing and applying corresponding standards and data infrastructures is therefore a major goal not only for health research but also for the public-health sector

in general, as has been impressively demonstrated during the current COVID-19 pandemic.

In the majority of cases, these types of health data are hosted in distributed local-data infrastructures within institutions, such as in hospitals, public-health authorities, and research institutions. Information on these data is often fractured in local repositories and websites or frequently even in publications.

It is therefore vital to enhance the findability and public availability of these distributed resources through the publication of GDPR-compliant rich metadata of health studies, including applied study instruments, questionnaires, data catalogues, and protocols. It is important to support the publication of health data by

1. developing and establishing domain-specific publication guidelines for health data;
2. enforcing the consistent application of existing health-data standards and metadata standards (including standard terminologies, ontologies, and minimal information standards for the health domain);
3. establishing new standards for health data and their metadata, if needed, in close collaboration with domain-specific standardization initiatives and committees of standards-developing organizations (SDOs), such as the ISO, the CEN, and their national counterparts;
4. providing incentives and support for publishing rich information about health data, including instruments, access information, patient-consent forms, and further study-related information;

Der Großteil des diesjährigen Berichts ist COVID-19-Aktivitäten gewidmet, außerdem unserem HITS Lab-Projekt DeepCu-rate und NFDI4Health, der Nationalen Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten. Zuvor jedoch fassen wir die Ziele der **Scientific Databases and Visualization (SDBV)** Gruppe zusammen und stellen die wichtigsten Entwicklungen vor.

Die Gruppe befasst sich mit Diensten über FAIR Daten. Die Abkürzung FAIR bezeichnet Findable, Accessible, Interoperable und Reusable Daten, also auffindbar, zugreifbar, interoperabel und nachnutzbare Daten. Mittels des SEEK Systems helfen wir Nutzern, ihre eigenen Daten FAIR zu machen. Mit SABIO-RK stellen professionelle Kuratorinnen FAIR-Daten über enzymkatalysierte Reaktionen bereit, extrahiert aus der Literatur. Der Sinn von FAIR ist die "Discoverability", also, dass sich Rechner in den Daten zurechtfinden können. Idealerweise sollte ein smartes Datensammelprogramm FAIRe Daten als solche erkennen, und bereit sein, die Daten zu ernten. Diese Ideen umsetzen zu wollen, impliziert Standardisierung. Partizipation in Standardisierungs-Initiativen wie COMBINE, STRENDA und ISO TC 276 Biotechnology, ist ein signifikanter Teil unserer Arbeit. Wichtig für unsere Arbeit ist die enge Zusammenarbeit von Biolog/-innen und Biochemiker/-innen, die mit Informatiker/-innen zusammenarbeiten.

Ein Großteil unserer Arbeit wird durch Projektförderung ermöglicht. Wir haben 3 Arten von Projekten: (i) Projekte mit biologischem/biomedizinischem Fokus, in denen SDBV ein Infrastruktur-Partner ist, wie z.B. LiSyM; (ii) kollaborative Infrastrukturprojekte mit SDBV, wie z.B. FAIRDOM und de.NBI; (iii) Forschungsprojekte wie DeepCurate und PoLiMeR.

Im letzten Jahr endete ein Projekt, das zusammen mit seinen Vorgängern SysMO-DB I und SysMO-DB II das vorige Jahrzehnt unserer Arbeit bestimmt hat: FAIRDOM, das Projekt über FAIR Daten, Operationen und Modelle. Dieses Projekt und von ihm abgeleitete Projekte waren hier oft Berichtsgegenstand. Auch in diesem Jahr sind Resultate aus FAIRDOM in zwei Projekten enthalten, die in diesem Jahr begonnen wurden, ELIXIR CONVERGE und NFDI4Health.

Wir hoffen, dass unsere Beiträge zu NFDI4Health und NFDI4Biodiversity, ELIXIR CONVERGE und anderen Projekten die FAIRDOM-Arbeiten weiter entwickeln werden.

5. supporting the harmonization of common-core (meta-)datasets. Funding and incentives for data collectors as well as corresponding funding for data stewards and data-management infrastructures is needed, especially for data harmonization and for past studies and their data (and to extract these data from the scientific literature).

Developed in the SDBV group with partners at the University of Manchester in the UK (Prof. Dr. Carole Goble's team), the SEEK platform will play a central role as a "hub" for data bundling, especially for metadata from various studies that are collected and sorted in NFDI4Health. The group will adapt the SEEK software to meet these requirements along with project partners in Leipzig, Göttingen, and

Greifswald. Together with project partner ZB Med in Cologne, HITS researchers will also develop a search portal that will access these data in order to make it easier for users to find them.

Martin Golebiewski will assume leadership of the entire NFDI4Health work package (task area) "Standards for FAIR Data" (TA2). Contributors to the work package will adapt and harmonize relevant data standards and thereby lay the foundation for bundling and structuring, which will facilitate finding and comparing the collected personal-health data. Martin has been involved in standardization committees at the International Organization for Standardization (ISO) for several years and leads the ISO/TC 276/WG 5, which comprises around 150 experts from all over the world

processing and integration in the life sciences. He is also involved in scientific-standardization initiatives, such as COMBINE (Computational Modeling in Biology Network) and EU-STANDS4PM (a European standardization framework for data integration and data-driven in silico models for personalized medicine). Our goal is to quickly and purposefully adapt standards to the requirements of the data in NFDI4Health and – if necessary – to initiate the development of new standards.

In support of the establishment and promotion of all research-data infrastructures, both the German Federal Government and individual states together intend to provide up to EUR 90 million annually until at least 2028 with the intention of creating a long-term sustained platform that supports research.

# 2 Research

## 2.12 Theory and Observations of Stars (TOS)



### Group Leader

Prof. Dr. Ir. Saskia Hekker (since September 2020)

### Staff member

Daria Mokrytska (since September 2020)

### Student

Julian Schlecker (since 14 December 2020)



Stars are an important source of electromagnetic radiation in the universe that allow for the study of many phenomena, ranging from distant galaxies to the interstellar medium and extra-solar planets. However, due to their opacity, "at first sight it would seem that the deep interior of the Sun and stars is less accessible to scientific investigation than any other region of the universe" (Sir Arthur Eddington, 1926). Now, however, through modern mathematical techniques and high-quality data, it has become possible to probe and study the internal stellar structure directly through global stellar oscillations via a method known as asteroseismology.

Asteroseismology uses similar techniques to helioseismology as carried out on our closest star, the Sun, to study the structure of other stars. The properties of waves are used to trace the internal conditions of these stars. Oscillations that impact upon the whole star reveal information that is hidden by the opaque surface. This asteroseismic information from the

CoRoT-, Kepler-, K2-, TESS-, SONG-, and Plato space observatories – combined with astrometric observations from Gaia, spectroscopic data from the SDSS-V APOGEE, interferometry, photometry, and state-of-the-art stellar models, such as MESA – provides insights into the stellar structure and physical processes that take place in stars.

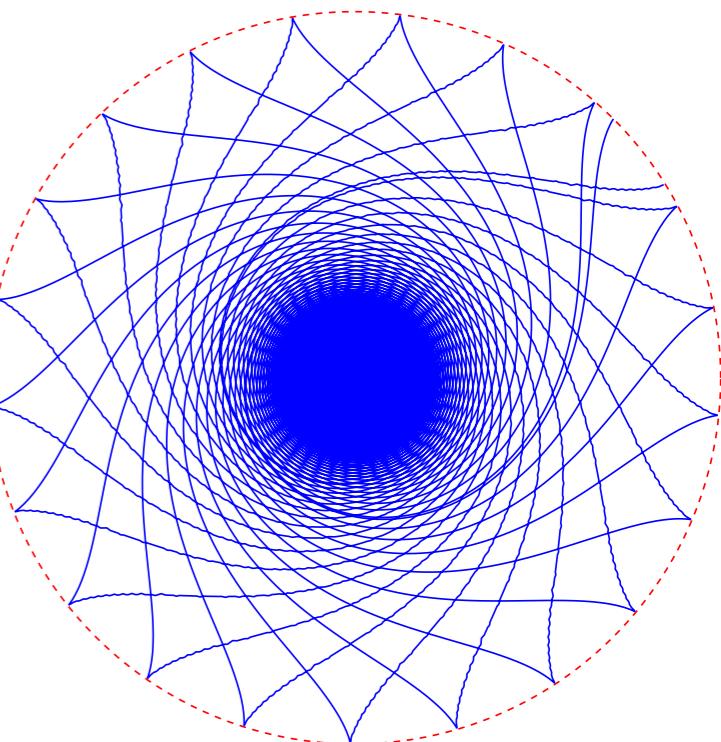
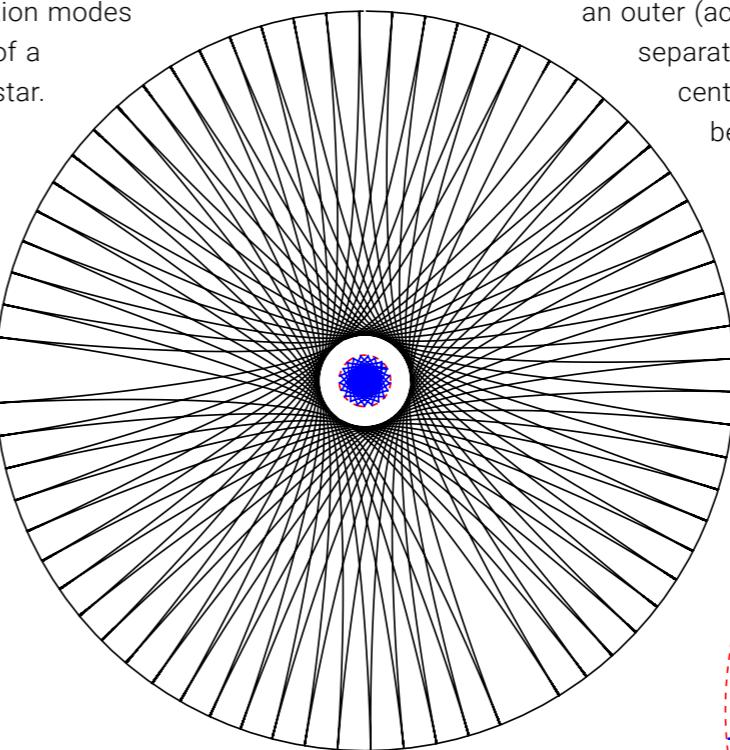
Understanding these physical processes within stars and how they change as a function of stellar evolution is the ultimate goal of the Theory and Observations of Stars (TOS) group at HITS, which was established in 2020. Our focus lies in – but is not limited to – low-mass main-sequence stars, subgiants, and red giants. These stars are interesting because they go through a series of internal structure changes. Furthermore, they are potential hosts of planets and serve as standard candles for galactic studies (core-helium-burning red giants). Exoplanet studies as well as galactic archaeology will hence also benefit from an increased understanding of these stars.

### Background

In the TOS group, we focus on stars with oscillations similar to those present in the Sun. These so-called solar-like oscillations are low-amplitude oscillations that are stochastically excited through turbulence in the near-surface convection layer of a star. The oscillations are sound waves and are expected to be present in all stars with convective outer layers. A convective envelope is typically present in low-mass main-sequence stars, subgiants, and red giants with surface temperatures below  $\sim 6,700$  °K. The stellar structure is imprinted in the global oscillation modes of a star.

An oscillation mode is uniquely determined by the properties of the matter through which it travels and is described by its frequency (or period) and mode identification – that is, the radial order (the number of nodal lines in the radial direction), spherical degree (the number of nodal lines on the surface), and azimuthal order (the number of nodal lines that cross the spin axis). In evolved, so-called red-giant stars, the dipole (with a spherical degree of 1) modes have sensitivity to both the deep interior and the outer layers – that is, the oscillations resonate in an inner (gravity) and an outer (acoustic) cavity separated by an evanescent zone (the area between the cavities where oscillations

cannot propagate and decay exponentially; see Fig.54). The coupling between the two oscillating cavities and the phases of the waves in each cavity can be derived from the resulting mixed pressure–gravity oscillations and can provide information on the physical conditions in the evanescent region. Furthermore, the difference in period between pure gravity dipole modes with consecutive radial orders (so-called period spacing, which can be extracted from mixed dipole modes) provides a measure of the extent of the gravity mode cavity and thus of the properties of the stellar core. Determining these values and understanding the physical processes that take place in these deep parts of stars is one of the aims of the TOS group.



*Figure 54 left: Ray path of a dipole mode in the outer acoustic (sound) cavity in black and of a dipole mode with the same frequency in the inner gravity cavity in blue. The outer turning points of the gravity modes are indicated by the red-dashed circle. The empty white ring between the black and blue ray paths is the evanescent zone. Right: Zoom of the ray path in the gravity mode cavity. The ray paths were computed following Alexander Kosovichev (Kosovichev, *Advances in global and local helioseismology: an introductory review*, lecture notes in physics 832, 3, 2011, <https://arxiv.org/abs/1103.1707>) based on a stellar model computed with the MESA stellar-evolution code (Paxton et al., *Modules for Experiments in Stellar Astrophysics (MESA): Pulsating variable stars, rotation, convective boundaries, and energy conservation*, ApJS 243, 10, 2019 and references therein, <https://arxiv.org/abs/1903.01426>).*

## Ongoing work

### Extraction of oscillation features

To extract the oscillations, we work in Fourier space. In this space, stellar oscillations appear as peaks. Due to their stochastic driving mechanism, solar-like oscillators have peaks with a width depending on their lifetimes. Identifying these peaks from the noise is therefore non-trivial. We developed a peak-detection method based on a Mexican-hat wavelet that automatically and reliably determines and characterizes the oscillation signals in terms of frequencies, amplitudes, and lifetimes (see Garcia Saravia Ortiz de Montellano, Hekker, Themeßl, Automated asteroseismic peak detections, MNRAS 476, 1470, 2018, <https://academic.oup.com/mnras/article/476/2/1470/4832523>). We are subsequently in the process of developing ways to characterize the modes in terms of their radial order, spherical degree, and azimuthal order. TACO (Tools for the Automated Characterization of Oscillations) – a full code that we plan to release soon – is in its final stages of development.

### Rotation

Given that the interior and exterior of stars rotate, mixed dipole modes always carry a signature of that rotation. The dipole modes are split into several components depending on the orientation of the star. If the rotation rate is high enough, this splitting can be observed. A significant part of the rotationally split mixed dipole modes – even of the most pressure-dominated ones – is sensitive to the core (Beck et al., Fast core rotation in red-giant stars as revealed by gravity-dominated mixed modes, Nature 481, 55, 2012

<https://www.nature.com/articles/nature10612?page=4>), and the core rotation rate therefore can be determined from the mixed dipole modes (Mosser et al., Spin down of the core rotation in red giants, A&A 548, A10, 2012, [https://www.aanda.org/articles/aa/full\\_html/2012/12/aa20106-12/aa20106-12.html](https://www.aanda.org/articles/aa/full_html/2012/12/aa20106-12/aa20106-12.html); Gehan et al., Core rotation braking on the red giant branch for various mass ranges, A&A 616, A24, 2018, <https://www.aanda.org/articles/aa/pdf/2018/08/aa32822-18.pdf>). However, to obtain the radial rotation profile, the envelope rotation rate is an essential input. Currently, the envelope-rotation rate has only been derived for about 20 red giants (Aerts, Mathis, Rogers, Angular momentum transport in stellar interiors, ARAA 57, 35, 2019, and references therein, <https://www.annualreviews.org/doi/abs/10.1146/annurev-astro-091918-104359>).

We are in the process of obtaining asteroseismic rotation measurements for a large set of hundreds of stars in order to build a statistical sample that can be used to study the angular momentum-transport mechanisms in stars. Furthermore, we aim to perform tests to validate the robustness of the results obtained from rotational inversions of the radial rotation profile in stars. More specifically, we intend to obtain these rotational inversion results for both lithium-rich stars and stars that are not rich in lithium. According to standard stellar-evolution models, lithium is depleted in evolved stars. However, lithium can be observed at enhanced values in a subsample of stars. Nearly all scenarios for

enhancing lithium include mixing, which could be a result of rotation. With this project, we aim to unravel the impact of rotation on the abundance of lithium.

### Red-giant bump

In addition to these ongoing projects, we aim to further our understanding of what happens at the red-giant bump. The red-giant bump is a short phase in stellar evolution in which changes in surface temperature and brightness are retracted before the stars continue on their paths. This feature is well known in models and observations; however, the physical mechanism behind it remains unknown, and the location of the bump in the models and observations is not the same. Last year, we published a paper (Hekker et al., Mirror principle and the red-giant bump: the battle of entropy in low-mass stars, MNRAS 492, 5940, 2020, <https://academic.oup.com/mnras/article/492/4/5940/5709942>), in which we proposed a physical mechanism that could cause this feature. We now plan to check whether the developed toy model indeed works and whether it can lead to a better understanding of the physics that is missing in the models.

## Projects beginning in 2021

### SFB 881 "The Milky Way System"

Beginning in 2021, the TOS group will participate in SFB 881 "The Milky Way System," in which we intend to use asteroseismology to obtain accurate masses, radii, ages, and distances as well as cluster memberships of individual cluster stars and to use this information to

improve the determined age and distance of the cluster when viewed as a single entity. At the same time, we aim to determine the present-day helium abundances for a subsample of stars. This latter determination will provide important constraints on our knowledge of the primordial helium abundance of the environment in which these stars were formed.

### ERC Consolidator Grant "Dipolar-Sound"

As of October 2021, the TOS group will also be funded by the ERC Consolidator Grant "DipolarSound." Within the framework of this grant, we will study stars that display different morphologies in their

Fourier spectrum. We aim to answer the following questions:

- What are the physical differences in the structures / conditions in red giants that lead to different oscillation spectra?
- What is the cause of the different structures / conditions in these stars?

The goal of the DipolarSound proposal is to unravel the physical conditions and processes at play in red giants using mixed dipole oscillation modes and to better understand the underlying physical origins of the different oscillation spectra observed in red giants.

Sterne sind eine wichtige Quelle elektromagnetischer Strahlung im Universum, mit der viele Phänomene untersucht werden können, von fernen Galaxien über das interstellare Medium bis hin zu Exoplaneten. Aufgrund ihrer Undurchsichtigkeit wurde jedoch einmal gesagt, dass „auf den ersten Blick das tiefste Innere der Sonne und der Sterne für wissenschaftliche Untersuchungen weniger zugänglich zu sein scheint, als jede andere Region des Universums“ (Sir Arthur Eddington, 1926). Durch moderne mathematische Methoden und die Menge und Qualität verfügbarer Daten ist es nun jedoch möglich geworden, die innere Sternstruktur direkt durch Sternschwingungen zu erforschen: eine Methode, die als Asteroseismologie bekannt ist.

Die Asteroseismologie verwendet ähnliche Techniken wie die Helioseismologie, die an unserem nächstgelegenen Stern, der Sonne, durchgeführt wird, um die Struktur anderer Sterne zu untersuchen. Hierzu werden die Eigenschaften von Wellen verwendet, um Rückschlüsse auf die innere Beschaffenheit von Sternen zu ziehen. Schwingungen, die auf den ganzen Stern einwirken, enthüllen so Informationen, die durch die undurchsichtige Oberfläche normalerweise verborgen sind. Diese asteroseismischen Informationen der Weltraumobservatorien wie CoRoT, Kepler, K2, TESS, SONG und Plato kombiniert mit astrometrischen Beobachtungen von Gaia, spektroskopischen Daten von SDSS-V APOGEE, Interferometrie, Photometrie und hochmodernen Sternmodellen wie MESA, geben Einblicke in die Sternstruktur und die physikalischen Prozesse, die in Sternen ablaufen.

Das Ziel der **Theory and Observations of Stars (TOS)** Gruppe am HITS, die 2020 eingerichtet wurde, ist die Untersuchung dieser physikalischen Prozesse, die in Sternen ablaufen, und wie sich diese in Abhängigkeit von der Sternentwicklung verändern. Die Gruppe konzentriert sich hierbei unter anderem auf sogenannte Hauptreihen-Sterne geringer Masse, „Unterriesen“ und rote Riesensterne. Diese Sterne sind deshalb interessant, weil sich ihre innere Struktur schnell ändert. Da sie potenziell von Planeten umgeben und kosmologische „Standardkerzen“ für Galaxienstudien sind, können sowohl die Exoplanetenforschung als auch die Galaxien-Archäologie vom wachsenden Verständnis dieser Sterne profitieren.

# 3 Centralized Services



## Group Leader

Dr. Gesa Schönberger

## Staff members

Christina Blach (office)  
Frauke Bley (human resources)  
Christina Bölk-Krosta (controlling)  
Benedicta Frech (office)  
Ingrid Kräling (controlling)  
Thomas Rasem (controlling)  
Rebekka Riehl (human resources and Assistant to the Managing Director)  
Stefanie Szymorek (human resources)  
Irina Zaichenko (accounting)

## Student

Lilly Börstler (until July 2020)

The HITS Lab program was established at about the same time. It provides additional funding for interdisciplinary group work on unusual research questions. We currently have two preliminary projects running that receive third-party funding and another two that are internally funded and began in 2020 (see Chapters 2.4, 2.6, 2.9, 2.11). In addition to its day-to-day business, the HR team spent almost the entire year working on selecting and introducing an HR software program. Personio software was finally chosen and implemented over the summer and began being used by all HITSters in September 2020.

The appointment of Saskia Hekker in the first half of the year and the launch of her TOS group in September 2020 are two additional and important major steps taken by the Institute (see Chapter 2.12). Supporting Saskia Hekker and Frauke Gräter in the application process for two ERC Consolidator Grants, which was met with success at the end of the year, was no less important. We celebrated another great milestone with Fabian Schneider, whose applications to both the DFG (Emmy Noether Program) and the EU (ERC Starting Grant) were successful. Fabian prepared with us for the start of his junior group, SET, in the last third of the year.

## 3.1 Administrative Services

The HITS administration supports the scientific groups in nearly all administrative tasks. It takes care of day-to-day operations at both HITS locations, manages human resources and accounting, clarifies legal issues, and assists the communications team in organizing events.

The corona pandemic was, at least in outward appearance, the main topic

of 2020, and this was also true for the administration at HITS. Beginning in March 2020, many procedures and processes were changed, and the administration team, like almost all other HITSters, switched from working in the office to working from remote locations without any major problems. All external and internal events and gatherings were initially cancelled, but many of these events could be rescheduled online after a short period of time. The cafeteria was closed from 12 March to 17 May and then again beginning on 21 December 2020, resulting in largely empty offices each time. A strong hygiene plan put rules in place for work for anyone who still wanted or had to come to HITS. This plan was regularly reviewed but could largely be maintained unaltered throughout the year. Thus, although many things at HITS were spatially separated, fortunately, the actual work itself could nevertheless proceed as usual in many cases.

After its official start at the beginning of 2020, the HITS administration was busy with the implementation of its new corporate design as well as with the first events in celebration of our 10-year anniversary (see chapter 5.3).

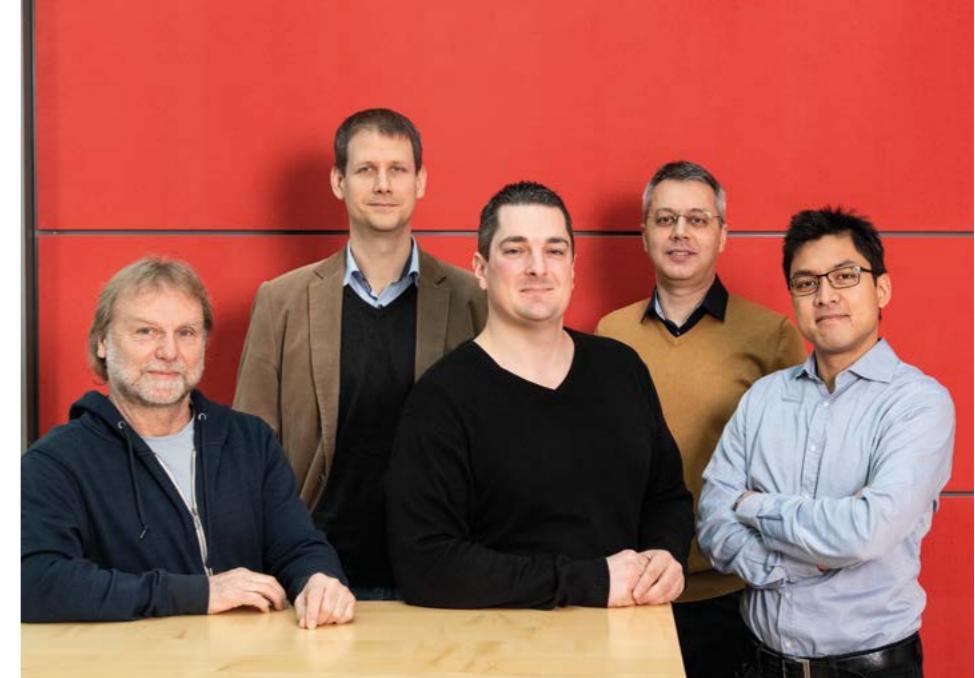
## 3.2 IT Infrastructure and Network

At the beginning of the year, we integrated a new underground fiber link in our network infrastructure, thereby finalizing a larger project that had begun in 2017. Both the Internet connection of the Institute and the connection between HITS and the Heidelberg University's Computing Centre have thus become fully redundant.

The beginning of the COVID-19 pandemic and the sudden switch to working from home in March did not catch us off guard. Indeed, for several years, HITS has had an extended system for remote work in place, including a secure VPN, a text-only- and a graphical login server, and many web services. Most HITSters were already equipped with the necessary laptops and accessories. The Internet connection of the Institute had sufficient free capacity to support sustained work from home as well as many videoconferences, which have since become an integral part of daily work at HITS. As a result, most

scientific activities have continued, even during the tight lockdown periods. The ITS group members have generally combined remote- and on-site work in order to maintain the hardware infrastructure while keeping a safe distance between each other.

The installation of the new HPC cluster at HITS, which had also been planned for the beginning of the year, had to be postponed to the summer due to large delays in the delivery of components. The old compute nodes, some of which dated back to 2010, were replaced by 150 new ones featuring Intel Cascade Lake CPUs and Infiniband technology of the newest generation for optimal scalability of our parallel codes.



## Group Leader

Dr. Ion Bogdan Costescu

## Staff members

Dr. Bernd Doser (Senior Software Developer)  
Norbert Rabes (System Administrator)  
Andreas Ulrich (System Administrator)  
Taufan Zimmer (System Administrator)

## Student

Julian Aris (July - December 2020)

The large BeeGFS storage system at HITS was also replaced, and the usable capacity was increased to approximately 1.6TB. The 150 compute nodes have access to this storage system over Infiniband as well as to the other BeeGFS storage system, which is located in the Heidelberg University Computing Centre, over 10/40G Ethernet. Sharing the storage systems increases efficiency by eliminating the need to copy/move data between the two clusters, and it also increases usability by employing a uniform naming scheme. This sharing will hopefully contribute to increased scientific output in the years to come.

# 4 Communication and Outreach



HITS Communications team in 2020 (f.l.t.r.): Alexandre Golovin, Peter Saueressig, Isabel Lacurie, Angela Michel.

## Head of Communications

Dr. Peter Saueressig

## Staff members

Isabel Lacurie

Angela Michel

## Student

Alexandr Golovin

The HITS communications team is the Institute's central hub for external and internal communications. Its main tasks are to raise the profile of HITS by coordinating media relations, digital and social-media communications, the Institute's publications, and design and branding as well as by organizing events for the scientific community,

We had been looking forward to 2020 with great anticipation because on 1 January 2020, HITS celebrated its 10th birthday. Therefore, we invited distinguished guests to our anniversary reception in January for the beginning of the many planned special events and activities (see Chapter 5.3). The reception was complemented by a one-page newspaper portrait in the local "Rhein-

Neckar-Zeitung" and an interview on the German radio station SWR 2 with HITS researchers Rebecca Wade (MCM), Sebastian Lerch (CST), and Fabian Schneider (PSO). In February, we organized the ZoRa Workshop at HITS, a project funded by the Klaus Tschira Foundation (see Chapter 5.1.3). Throughout the course of the year, we had planned to organize an anniversary open-house event and a

series of public talks in the MAINS venue in Heidelberg. At the first event in February, Frauke Gräter and Isabel Martin presented "crash test for proteins." However, in March, all our plans came crashing down due to corona. Many events had to be canceled, and others were shifted to the digital sphere. As our lab is the computer, we had fewer problems coping than



*HITS on the air: Radio journalist Eberhard Reuß interviews Rebecca Wade (MCM) for the broadcast on the Institute's anniversary in January 2020.*

other institutes, but we all missed seeing one another in person. To compensate for this loss, we immediately intensified our internal communications by issuing regular emails to all HITSters to inform them of the current status of the pandemic, and we included some lab news

management and standardization to the search for suitable medicine and COVID-19 forecasts – projects no one would have envisioned in January. "Non-corona"-related research was also successful and included new, "explosive" findings on dying stars, a study on collagen with an impact on material research and biomedicine, and an easy-to-use open-source online platform for biomedical image segmentation.

In terms of the HITS workforce, we literally reached for the stars: In September, asteroseismologist Saskia Hekker joined HITS as TOS Group Leader (see Chapter 2.12.). In December, she won an ERC Consolidator Grant for her research on the sound of the stars. At the same time,



*"Crash test for proteins": HITS researchers Frauke Gräter (right) and Isabel Martin (left) presented their research in the MAINS Heidelberg, moderated by science journalist Pia Grzesiak.*

about special achievements and interesting topics. As there were many such topics, there was always enough "fodder" for us to broadcast.

## Scientific excellence: Fighting COVID and reaching for the stars

In terms of scientific success, the HITS communications team was pleased to announce a great deal of good news to the public. Our researchers participated in COVID-19 projects ranging from data



MBM Group Leader Frauke Gräter won an ERC Consolidator Grant for her project on radicals in collagen (see Chapter 2.7).

Additionally, we were granted even more funding from the ERC: Kai Polsterer's AIN group participated in a project that received an ERC Synergy Grant for measuring the true scale of the Universe. Moreover, astrophysicist Fabian Schneider received an ERC Starting Grant for his project on the evolution of stars and began his own research group at HITS in January 2021. Speaking of

stars: For five years running, Clarivate Analytics listed computer scientist Alexandros Stamatakis as one of the most-cited researchers in his field worldwide.

Furthermore, we were able to present a plethora of research highlights from the last ten years in our newly established blog "Via Data," edited by Angela Michel on the "Scilogs" platform. Throughout the course of the year, several researchers joined the team of authors and posted about their experiences and findings in easily understandable language. Inspired by the positive response, we will continue with the "Via Data" blog beyond the anniversary year.

## Media relations: "Journalist in Residence" program

We firmly believe that an important prerequisite for successful science communication is the development of reliable and sustainable journalistic contacts. The "Journalist in Residence" program thus represents an important project for HITS. The program is geared toward science journalists and offers them a paid sojourn at HITS. During their stay, journalists can learn more about data-driven science and get to know

researchers and new research topics without the pressure of the “daily grind.” In the most-recent opening, candidates from six continents applied.

Canadian science journalist and author Siobhan Roberts was chosen to be the 2020 “Journalist in Residence.” Roberts has worked as a freelance journalist with a focus on mathematics and science since 2001. She writes regularly for The New York Times’ “Science Times” and has contributed to The New Yorker’s science and tech blog, “Elements,” as well as to The Walrus, Quanta, and The Guardian, among other publications. Moreover, Roberts is the author of two biographies on mathematicians: “King of Infinite Space,” on Donald Coxeter, and “Genius at Play,” on John Horton Conway. She has earned multiple awards for her work, including the Euler Book Prize from the Mathematical Association of America.

Siobhan Roberts arrived at HITS in



Siobhan Roberts, HITS Journalist in Residence 2020.

mid-September 2020. In spite of the Corona crisis, she was able to meet with HITS researchers from different



*“Run beyond the limits!”: The HITS running team at the first virtual NCT run in June completed a combined total distance of 427 kilometers and raised more than EUR 10,000 for cancer research.*

groups on several occasions, both online and “in real life.” Roberts’s goal was to increase her fluency in data-driven research and to explore opportunities for data journalism. She also worked on her current book project, a biography on Swiss-American/Canadian mathematical logician and group theorist Verena Huber-Dyson. Roberts offered HITS re-

searchers an online writing workshop in November, which was fully booked. Due to the pandemic, she was only able to visit the numerous university- and extramural research institutes in Heidelberg on rare occasions; however, Roberts gave an online public talk entitled “Embracing the Uncertainties” in January 2021 that marked the end of her stay (see the video of her talk



Kai Polsterer delivering his keynote address at the virtual ADASS conference in November 2020 via video stream.

on the HITS YouTube channel). “HITS is that rare place where you can buckle down and make progress on a big project while at once expanding your reportorial horizons in all sorts of new directions. And the hilly forest provides the perfect escape to contemplate and process it all and map a path forward,” stated Roberts, summing up her stay at the Institute. The next HITS “Journalist in Residence” will be radio journalist Carl Smith (Sydney, Australia) in the second half of 2021.

### Events: Separated by distance, but united in spirit

In addition to many digital events, such as the “virtual lab visit” for young researchers during the Heidelberg Laureate Forum (see Chapter 6) and the live streaming of Kai Polsterer’s keynote address from HITS as part of the ADASS conference, some “in-person” events also took place. The annual run organized by the National Center for Tumor Diseases (NCT) in Heidelberg is one of the most-popular charity events in the region.

In 2020, the run had to be reorganized as a virtual three-day event. From 26–28 June, almost 9,000 participants ran a total of 120,000 kilometers either individually or in small groups following a route of their own choosing. Among them was the HITS team of 30 runners, who were especially motivated: On the occasion of its 25th anniversary, the Klaus Tschira Foundation had decided to sponsor each staff kilometer with 25 euros. And the running team did not need to be told twice: In keeping with the team slogan “Run beyond the limits!” they completed a combined distance of 427 kilometers and thus raised a total of EUR 10,675 for cancer research. The best moments of this first virtual NCT Run are preserved for posterity on YouTube.

Another event took place from 20 September to 10 October 2020: the nationwide “Stadtradeln” (“City Cycling Challenge”) competition. The competition was designed to show that we can easily navigate through daily life by bike and thereby contribute to saving the environment. 133

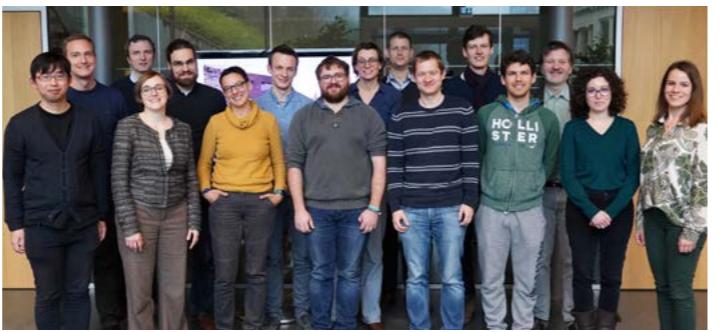
teams participated in Heidelberg alone, bicycling more than 200,000 kilometers in total. The “HITster” team ranked 25th in the citywide competition and contributed over 2,000 kilometers. Thanks to their efforts, the team managed to save over 300 kilograms of CO<sub>2</sub> emissions while having fun and staying healthy.

# 5 Events

## 5.1 Conferences, Workshops & Courses

### 5.1.1 Emulator Day

**27 January 2020**  
**Studio Villa Bosch/HITS,  
Heidelberg, Germany**



What does research on astrophysics, molecular biology, and weather prediction have in common? The relevant physical processes act – and need to be modeled – on a wide variety of spatial and temporal scales. Up to now, these scales have typically been addressed with ad-hoc models and so-called "sub-grid-scale" physics. However, using complimentary methods from computational statistics and machine learning appears promising and could lead to faster, more-efficient simulations. On the interdisciplinary "Emulator Day," organized by HITS group leaders Tilmann Gneiting (Computational Statistics), Frauke Gräter (Molecular Biomechanics), and Friedrich Röpke (Physics of Stellar Objects), scientists met at the Studio Villa Bosch and at HITS to discuss emulators – that is, judicious representations of complex, physics-based computer code via simpler and much-faster surrogate models.

The event began with a colloquium talk by Corinna Hoose (Karlsruhe Institute of Technology KIT, Germany) on the "Simulation of Deep Convective Clouds under Various Meteorological and Microphysical Impacts." Afterward, Tilmann Gneiting presented an overview on emulators from a statistical perspective. Constanze Wellmann (Heidelberg University, Germany) provided insights on her use of "Statistical Emulation for Sensitivity Studies of Deep Convective Clouds," and Takahiro Nishimichi (Kyoto University, Japan) gave a talk about "Emulating Halo Statistics for Large-scale Structure Cosmology." The "Emulator Day" also marked the kick-off of a new interdisciplinary project called "Emulation in Simulation," a HITS Lab project led by group leaders Tilmann Gneiting, Frauke Gräter, and Friedrich Röpke (see also Chapter 2.4).

### 5.1.2 EU-STANDS4PM Workshop "Using patient-derived data for in silico modeling in personalized medicine"

**4 February 2020**  
**HITS, Heidelberg, Germany**

EU-STANDS4PM is a European initiative funded by the Horizon2020 Framework Programme of the European Commission with the aim of developing and establishing a European standardization framework for data integration and data-driven in silico models for personalized medicine. In an EU-STANDS4PM workshop organized by Martin Golebiewski, about 25 experts in data- and model standardization, data collection, and model simulation in personalized medicine met in Heidelberg. The attendees originated from 8 different countries – mainly from Germany and other parts of Europe but also from New Zealand – and represented the fields of academic and industrial research as well as clinical applications. Many of the attendees presented their use cases on using patient-derived data for in silico modeling in short presentations. To initiate a standardization roadmap for the use



of artificial-intelligence (AI) technology and manual in silico modeling approaches for personalized medicine, HITSter Martin Golebiewski together with Heike Moser from DIN (German Institute for Standardization) presented possible options for standardization strategies. Moreover, Christina Kyriakopoulou, a representative from the European Commission's (EC) Health Directorate, attended the workshop and gave an overview on the EC's vision of the importance of standardization for in silico modeling in personalized medicine.



The workshop focused on the need for the standardization of patient-derived data for use in disease-related modeling in the context of personalized medicine as well as on the potential for artificial intelligence to utilize this patient-derived data and to support the modeling process. In addition to the gathering and discussion of use cases, the major aims of the workshop were also to compare state-of-the-art modeling approaches based on patient-derived data, to discuss the potential for artificial-intelligence (AI) methods to utilize patient-derived data for modeling, as well as to analyze standardization gaps in the use of AI technologies and manual modeling approaches in the context of in silico models for personalized medicine.

### 5.1.3 ZOrA Workshop

**28 February 2020**  
**HITS, Heidelberg, Germany**



On 28 February 2020, HITS hosted a workshop within the framework of the "Zukunfts-Orientierungs-Akademie" ("Future-Orientation Academy"; short: ZOrA). The academy is a project by the PH Heidelberg and Heidelberg University and is funded by the Klaus Tschira Stiftung. HITS – among other Heidelberg institutions – is a scientific partner of the program. ZOrA offers female high-school students the possibility to visit different institutes and organizations in Heidelberg in order to find out about career options in IT and science. HITS researchers Saskia Haupt (DMQ), Sabrina Gronow (PSO), and Mareike Pfeil (GRG) provided some valuable insights into their study paths as well as into their current research and career goals. Afterwards, the girls had the possibility to ask the three scientists all the questions they had always wanted to ask about becoming a scientist.



### 5.1.4 Workshop "FAIR Data Infrastructures for Biomedical Communities"

**15 October 2020**  
**online**



This online workshop with more than 80 attendees from Germany and all over Europe introduced several European and national activities geared toward supporting the biomedical research community with FAIR (Findable, Accessible, Interoperable, and Reusable) data management. By exchanging experiences from across the different initiatives, common concepts and a joint strategy for FAIR data and corresponding standards were developed. HITSter Martin Golebiewski organized the workshop together with partners in Göttingen and Leipzig as a satellite of the 65th annual conference of the German Association for Medical Informatics, Biometry and Epidemiology (GMDS). Martin leads the hosting GMDS working group "FAIR data infrastructures for biomedical informatics," which was formed in early 2019 (<https://bit.ly/2N6rCq1>). The group bundles infrastructure activities for health-related FAIR data in the fields of biomedical research, clinical-data management, and medical informatics in Germany. It aims to provide a platform for information exchange on the best practices for and the implementation of FAIR data management in these fields. Keynotes were presented by Niklas Blomberg – Director of the European Bioinformatics Infrastructure ELIXIR – and Carole Goble from the University of Manchester (UK). Sylvia Thun from the Berlin Institute of Health at Charité spoke about the highly topical "Standardization and interoperability of study data from COVID-19 clinical and public-health research." Several initiatives for FAIR data from the domain presented their activities, including FAIRsharing, FAIR4Health, FAIRness for FHIR, the European Open Science Cloud for the life sciences (EOSC-Life), and FAIRDOM. The workshop also aimed to collect requirements for the newly formed German National Research Data Infrastructure (NFDI). The coordinators of both the German Human Genome-Phenome Archive (GHGA) and the National Research Data Infrastructure for Personal Health Data (nfdi4health) gave an update on these parts of NFDI.

## 5.1.5 Astrophysics winter workshop

**14–18 December 2020**  
online

Due to the COVID-19 pandemic, the annual winter workshop organized by the Physics of Stellar Objects (PSO) group had to take place as an online event. Despite its disadvantages in terms of discussions, the format allowed for more participants to take part in the "15th Würzburg workshop" than had done so in previous years and for the program to be extended to cover a full week (from 14–18 December). The international group of about 50 participants included scientists from Australia, China,

Sweden, the United Kingdom, Ireland, the United States, and several institutions in Germany and discussed topics on supernova research, binary stellar evolution, asteroseismology, and processes in stellar interiors.

The plenary session featured invited talks by Sherry Suyu (MPA Garching, Germany) on gravitationally lensed supernovae, by Zhengwei Liu (Yunnan Observatories of the Chinese Academy of Sciences) on the impact of supernova explosions on companion stars, by Saskia Hekker (HITS) on asteroseismology, by Johann Higl

(HITS) on compressible simulations of stellar oscillations, by Mark Magee (Trinity College Dublin, Ireland) on light curves of Type Ia supernovae, and by Patrick Ondratschek (MPS Göttingen, Germany) on magnetohydrodynamic simulations of common envelope phases in stellar binary systems. The plenary session was followed by topical sessions in which the participants presented their latest results, and projects for collaborations were discussed.

**Prof. Dr. Jens Meiler**



**Vanderbilt University, Informatics Center for Structural Biology, USA**

8 September 2020: Innovative Computational Methods for Protein Structure Prediction, Drug Discovery, and Therapeutic Design (Online)

**Prof. Dr. Claudia Draxl**



**Physics Department of the Humboldt-University of Berlin, Germany**

12 October 2020: Predicting properties of complex materials: challenges for modern ab initio theory (Online)

**Prof. Dr. Robert Best**



**National Institute of Diabetes and Digestive and Kidney Diseases, Theoretical Biophysical Chemistry Section, USA**

23 November 2020: Molecular Simulations of Intrinsically Disordered Proteins (Online)

## 5.2 HITS Colloquia

**Prof. Dr. Corinna Hoose**

**Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Germany**

27 January 2020: Simulation of deep convective clouds under various meteorological and microphysical impacts

**Dr. Wolfgang Maier**

**IBM R&D – Systems & Technology Group, Böblingen, Germany**

17 February 2020: Future Compute Paradigms

**Prof. Dr. Robert C. Williamson**

**Australian National University, Research School of Computer Science, Australia**

29 June 2020: The AI of Ethics (Online)



## 5.3 HITS anniversary reception

HITS was established on 1 January 2010 when EML Research gGmbH changed its name. In 2020, the institute celebrated its 10th anniversary. The first event to mark the occasion was the anniversary reception on 20 January 2020. HITS Managing Director Gesa Schönberger and Scientific Director Wolfgang Müller welcomed more than 80 guests to the Studio Villa Bosch: scientific and industrial partners, local Heidelberg politicians, science journalists, and friends.

Gesa Schönberger and Wolfgang Müller presented a review of the last 10 years and an outlook on upcoming events (unfortunately, most of them had to be canceled due to the pandemic), followed by remarks by Andreas Reuter, former Managing Director and member of the board of directors of the HITS-Stiftung.



Cutting the first slice: Wolfgang Müller and Gesa Schönberger.



10 years HITS: The fancy birthday cake.



The audience in the Carl Bosch Auditorium.



Kai Polsterer during his talk about AI in science.

Next, group leader Kai Polsterer (AIN) gave a vivid and lively talk about AI in science entitled "Wenn Maschinen lernen: Künstliche Intelligenz in der Wissenschaft" ("When Machines Learn: Artificial Intelligence in Science").

Afterward, during the informal reception, the HITS management cut the first slice of a fancy birthday cake as a symbolic start to the anniversary year.

# 6 Collaborations



Anna Wienhard and Andreas Reuter at the HLF closing ceremony (Photo: HLFF).

## Heidelberg Laureate Forum

The Heidelberg Laureate Forum (HLF) is a networking conference at which 200 carefully selected young researchers in mathematics and computer science spend a week interacting with the laureates of the disciplines, including recipients of the Abel Prize, the ACM A.M. Turing Award, the ACM Prize in Computing, the Fields Medal, and the Nevanlinna Prize. Established in 2013, the HLF is organized annually by the Heidelberg Laureate Forum Foundation (HLFF). The event was the product of a joint initiative between HITS and the Klaus Tschira Foundation. Since 2016, HITS has served as a scientific partner of the HLF along with Heidelberg University.

### Traversing separation: The Virtual HLF, 21–25 September 2020

The striking difference in 2020 compared with former HLF events was the obvious shift to the digital sphere. Branded with the motto "Virtual HLF – Traversing Separation," the Forum took place from 21–25

September 2020. It included a multifaceted scientific program and various interactive elements designed to facilitate exchange between participants. Tools such as the interactive event app and the Virtual Meeting Hub in VR were heavily employed by the young researchers throughout the week.

#### **Virtual lab visit: HITSters meet young researchers**

This program includes lab visits during which a group of young researchers come to Heidelberg institutes and companies to get a glimpse of the work done there. Upon assuming her new position, she proclaimed, "I look forward maintaining the tradition of excellence at the HLF and to shaping its future." Anna took on her new role on 1 October 2020.

HITS Scientific Director Wolfgang Müller led a virtual campus tour and explained the range of research performed at the Institute. Afterward, CME group leader Alexandros Stamatakis gave a remote talk on the "phylogenetic inference of SARS-CoV-2" from his vacation home in Crete – one of the advantages of a virtual event. More than 40 young

#### **Anna Wienhard as the new scientific chairperson of the HLFF**

The closing ceremony of the HLF saw a change in the conference committee: As planned, computer scientist and former HITS Managing Director Andreas Reuter announced that he was stepping down as

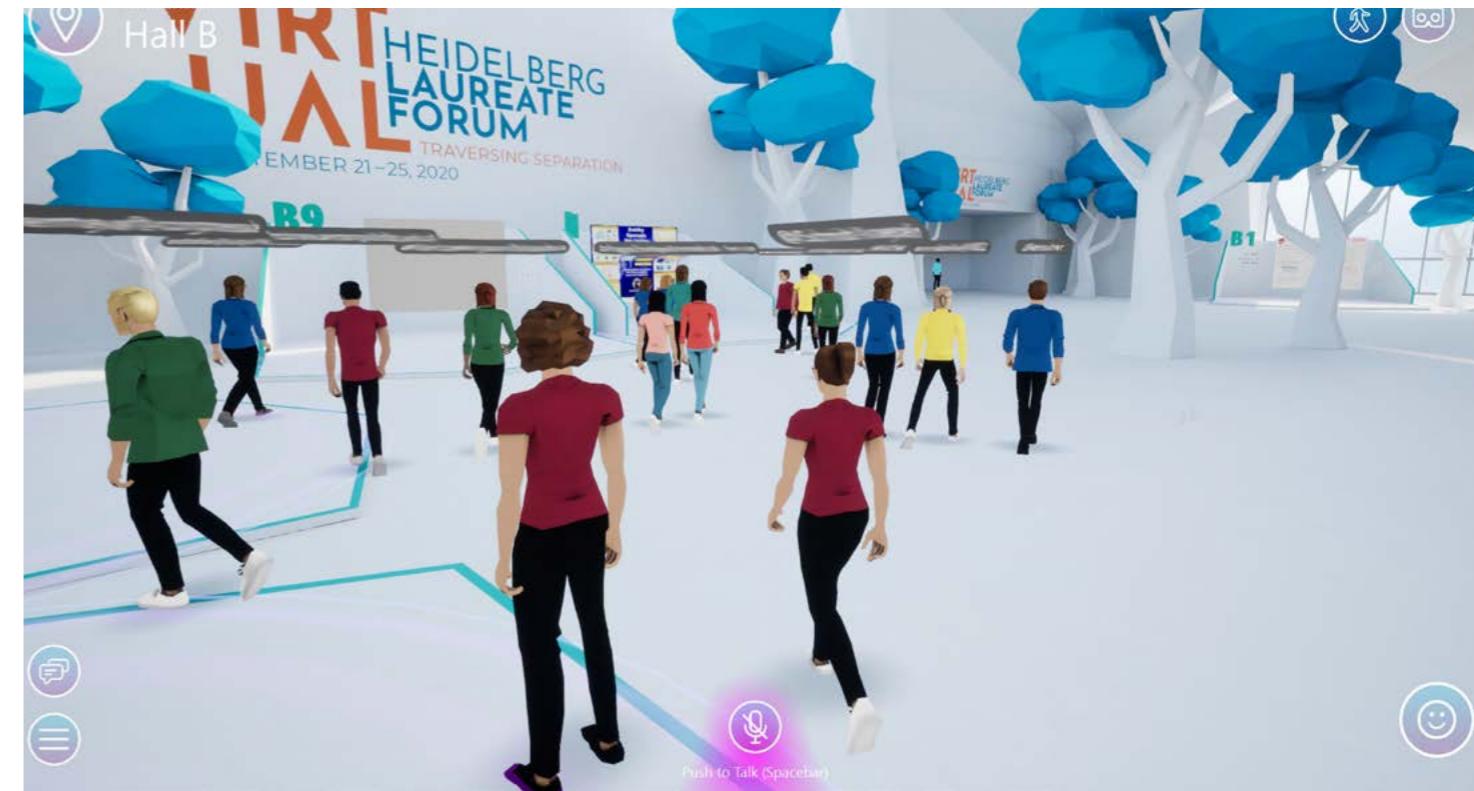
Scientific Chairperson of the HLFF, a position he had held since the Foundation was established in 2013. Andreas was essential in developing the scientific elements of each HLF program. He passed the proverbial torch to mathematician Anna Wienhard, GRG group leader at HITS and professor at Heidelberg University. Anna has been an active member of the HLF's Scientific Committee for several years and is an avid supporter of the HLF who has helped promote its guiding vision. Upon assuming her new position, she proclaimed,

"I look forward maintaining the

tradition of excellence at the HLF and to shaping its future."

Anna took on her new role on 1 October 2020.

researchers from four continents and in different time zones participated in the session and made extensive use of the Q&A to pose questions about what it is like to conduct research and work at HITS. The session went smoothly thanks to HITS Alumnus Volker Gaibler, now a member of the HLF staff, who moderated the event.



A get-together in times of physical distancing: HLF participants during the Virtual Meeting Hub (Picture: HLFF).

## Informatics4Life

Despite remarkable progress in the diagnosis and treatment of acute and chronic cardiovascular diseases in recent years, these illnesses remain the leading cause of death and hospitalization for all people worldwide. Informatics4life is a collaborative initiative founded by the Klaus Tschira Foundation that focuses on cardiovascular research and brings together experts in clinical research and computational methods. The initiative's patient-centric environment – a completely novel approach – is critical to translating research into clinical application. As Heidelberg is a "hot spot" for health research and medicine with internationally highly competitive institutes and researchers, it is the ideal location for this pioneering initiative.

The project represents an interdisciplinary alliance between cardiovascular physicians and computer scientists from Heidelberg University, University Hospital Heidelberg, and

HITS. It consists of several subprojects in which HITS researchers are involved as principal investigators and contributors.



#### **HITS involvement: From structure based-design to Machine Learning**

Rebecca Wade (MCM) is a Principal Investigator of the subproject "Structure-based design of peptide-based pharmaceuticals against striated muscle disorders" (see Chapter 2.8), Wolfgang Müller (SDBV) is Collaborating Principal Investigator of the subproject "Research data warehouse," and Vincent Heuveline (DMQ) is Principal Investigator of the subproject "Cognition and uncertainty quantification for numerical heart simulation" (see Chapter 2.5). Moreover, Vincent is also Co-Coordinator of Informatics4life in cooperation with Hugo Katus and Benjamin Meder (both from Heidelberg University Hospital).

Another subproject in which HITSters are involved is "ML-supported high-content screening of temporal and spatial cellular features" (ML = machine learning). This project has additionally been funded by the Klaus Tschira Foundation and – due to its similar focus – has recently merged with Informatics4 Life. Kai Polsterer (AIN) and Jan Plier work on this project and collaborate with Mathias Konstandin (Heidelberg University Hospital). Even though the scales are extremely different, morphological analyses of cell tissues are very similar to analyses of multi-band observations of galaxies – an observation that provides further support for Klaus Tschira's vision that interdisciplinary research is key to making progress in research (see Chapter 2.1).

# 7 Publications

Ballhausen A, Przybilla MJ, Jendrusch M, Haupt S, Pfaffendorf E, Seidler F, Witt J, Sanchez AH, Urban K, Draxlbauer M, Krausert S, Ahadova A, Kalteis MS, Pfuderer PL, Heid D, Stichel D, Gebert J, Bonsack M, Schott S, Bläker H, Seppälä T, Mecklin J, Broeke ST, Nielsen M, Heuveline V, Krzykalla J, Benner A, Riemer AB, Doeberitz MvK, Kloot M (2020). The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoediting during tumor evolution. *Nat Commun* 11(1), 4740

Barbera P, Czech L, Lutteropp S, Stamatakis A (2020). SCRAPP: A tool to assess the diversity of microbial samples from phylogenetic placements. *Mol Ecol Resour*, Vol. 21, Issue 1/ p. 340-349

Berger B-T, Amaral M, Kokh DB, Nunes-Alves A, Musil D, Heinrich T, Schröder M, Neil R, Wang J, Navratilova I, Bomke J, Elkins JM, Müller S, Frech M, Wade RC, Knapp S (2020). Structure-Kinetic-Relationship Reveals the Mechanism of Selectivity of FAK Inhibitors Over PYK2. *Cell Chem Biol* S2451-9456(21):00003-9

Bestenlehner JM, Crowther PA, Caballero-Nieves SM, Schneider FR, Simón-Díaz S, Brands SA, de Koter A, Gräfener G, Herrero A, Langer N, Lennon DJ, Maíz Apellániz J, Puls J, Vink JS (2020). The R136 star cluster dissected with Hubble Space Telescope/STIS - II. Physical properties of the most massive stars in R136. *Monthly Notices of the Royal Astronomical Society* 499(2):1918-1936

Bettsworth B, Stamatakis A (2020). RootDigger: a root placement program for phylogenetic trees. *bioRxiv* 2020.02.13.935304

Blacker S, Bastian NF, Bauswein A, Blaschke DB, Fischer T, Oertel M, Soulardis T, Typel S (2020). Constraining the onset density of the hadron-quark phase transition with gravitational-wave observations. *Phys. Rev. D* 102(12), 123023

Bläker H, Haupt S, Morak M, Holinski-Feder E, Arnold A, Horst D, Sieber-Frank J, Seidler F, von Winterfeld M, Alwers E, Chang-Claude J, Brenner H, Roth W, Engel C, Löffler M, Mösllein G, Schackert H, Weitz J, Perne C, Aretz S, Hüneburg R, Schmiegel W, Vangala D, Rahner N, Steinke-Lange V, Heuveline V, von Knebel Doeberitz M, Ahadova A, Hoffmeister M, Kloot M (2020). Age-dependent performance of BRAF mutation testing in Lynch syndrome diagnostics. *Int. J. Cancer* 147(10):2801-2810

Bodensteiner J, Sana H, Mahy L, Patrick LR, de Koter A, de Mink SE, Evans CJ, Götberg Y, Langer N, Lennon DJ, Schneider FRN, Tramper F (2020). The young massive SMC cluster NGC 330 seen by MUSE. I. Observations and stellar content. *A&A* 634:A51

Bowman DM, Burssens S, Simón-Díaz S, Edelmann PVF, Rogers TM, Horst L, Röpke FK, Aerts C (2020). Photometric detection of internal gravity waves in upper main-sequence stars. *A&A* 640:A36

Brehmer, J.R., Gneiting, T. Properization: constructing proper scoring rules via Bayes acts. *Ann Inst Stat Math* 72, 659–673 (2020)

Brenna C, Khan AUM, Picascia T, Sun Q, Heuveline V, Gretz N (2020). New technical approaches for 3D morphological imaging and quantification of measurements. *Anat Rec* 303(10):2702-2715

Brunak S, Collin CB, EU-STANDS4PM Consortium, Cathaoir KEÓ, Golebiewski M, Kirschner M, Kockum I, Moser H, Waltemath D (2020). Towards standardization guidelines for *in silico* approaches in personalized medicine. *Journal of Integrative Bioinformatics* 17(2-3)

Chai H, Zhao W, Eger S, Strube M (2020). Evaluation of Coreference Resolution Systems Under Adversarial Attacks. In Proceedings of the First Workshop on Computational Approaches to Discourse, Online, November, pp. 154-159

Diestelkoetter-Bachert P, Beck R, Reckmann I, Hellwig A, Garcia-Saez A, Zelman-Hopf M, Hanke A, Nunes-Alves A, Wade RC, Mayer MP, Wieland F (2020). Structural characterization of an Arf dimer interface: molecular mechanism of Arf-dependent membrane scission. *FEBS Lett*, 594(14):2240-2253

Doorenbos L, Cavuoti S, Brescia M, D'Isanto A, Longo G, (2020). Comparison of outlier detection methods on astronomical image data. *Intelligent Astrophysics*, Springer Nature 10.1007/978-3-030-65867-0

Edgar RC, Taylor J, Altman T, Barbera P, Meleshko D, Lin V, Lohr D, Novakovsky G, Al-Shayeb B, Banfield JF, Korobeynikov A, Chikhi R, Babaian A (2020). Petabase-scale sequence alignment catalyses viral discovery. *bioRxiv* 2020.08.07.241729

Elsworth Y, Themeßl N, Hekker S, Chaplin W (2020). A Layered Approach to Robust Determination of Asteroseismic Parameters. *Res. Notes AAS* 4(10):177

Evans C, Lennon D, Langer N, Almeida L, Bartlett E, Bastian N, Bestenlehner J, Britavskiy N, Castro N, Clark S, Crowther P, de Koter A, de Mink S, Dufton P, Fossati L, Garcia M, Gieles M, Gräfener G, Grin N, Hénault-Brunet V, Herrero A, Howarth I, Izzard R, Kalari V, Maíz Apellániz J, Markova N, Najarro F, Patrick L, Puls J, Ramírez-Agudelo O, Renzo M, Sabín-Sanjuán C, Sana H, Schneider F, Schootemeijer A, Simón-Díaz S, Smartt S, Taylor W, Tramper F, Van Loon J, Villaseñor J, Vink JS, Walborn N (2020). The VLT-FLAMES Tarantula Survey. *The Messenger*, 181:22-27

Franz F, Daday C, Gräter F (2020). Advances in molecular simulations of protein mechanical properties and function. *Current Opinion in Structural Biology* 61:132-138

Galvin TJ, Huynh MT, Norris RP, Wang XR, Hopkins E, Polsterer K, Ralph NO, O'Brien AN, Heald GH (2020). Cataloguing the radio-sky with unsupervised machine learning: a new approach for the SKA era. *Monthly Notices of the Royal Astronomical Society* 497(3):2730-2758

Gornik SG, Bergheim BG, Morel B, Stamatakis A, Foulkes NS, Guse A (2020). Photoreceptor diversification accompanies the evolution of Anthozoa. *Molecular Biology and Evolution*, msaa304

Gottschling M, Czech L, Mahé F, Adl S, Dunthorn M (2020). The windblown: possible explanations for dinophyte DNA in forest soils. *bioRxiv* 2020.08.07.242388

Gronow S, Collins C, Ohlmann ST, Pakmor R, Kromer M, Seitenzahl IR, Sim SA, Röpke FK (2020). SNe Ia from double detonations: Impact of core-shell mixing on the carbon ignition mechanism. *A&A* 635:A169

Haupt S, Gleim N, Ahadova A, Bläker H, von Knebel Doeberitz M, Kloot M, Heuveline V (2020). Computational model investigates the evolution of colonic crypts during Lynch syndrome carcinogenesis. *bioRxiv* 2020.12.29.424555

Haupt S, Zeilmann A, Ahadova A, von Knebel Doeberitz M, Kloot M, Heuveline V (2020). Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis using Dynamical Systems with Kronecker Structure. *bioRxiv* 2020.08.14.250175

Holderbach S, Adam L, Jayaram B, Wade RC, Mukherjee G (2020). RASPD+: Fast Protein-Ligand Binding Free Energy Prediction Using Simplified Physicochemical Features. *Frontiers in Molecular Biosciences* 7:393

Hon M, Bellinger EP, Hekker S, Stello D, Kuszlewicz JS (2020). Asteroseismic inference of subgiant evolutionary parameters with deep learning. *Monthly Notices of the Royal Astronomical Society* 499(2):2445-2461

Horst L, Edelmann PVF, Andrásy R, Röpke FK, Bowman DM, Aerts C, Ratnasingam RP (2020). Fully compressible simulations of waves and core convection in main-sequence stars. *A&A* 641:A18

Jeon S, Strube M (2020). Incremental Neural Lexical Coherence Modeling. In Proceedings of the 28th International Conference on Computational Linguistics (COLING), Online, December, pp. 6752-6758

Jeon S, Strube M (2020). Centering-based Neural Coherence Modeling with Hierarchical Discourse Segments. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, November , pp. 7458-7472

Keating SM, Waltemath D, König M, Zhang F, Dräger A, Chaouiya C, Bergmann FT, Finney A, Gillespie CS, Helikar T, Hoops S, Malik-Sheriff RS, Moodie SL, Moraru II, Myers CJ, Naldi A, Olivier BG, Sahle S, Schaff JC, Smith LP, Swat MJ, Thieffry D, Watanabe L, Wilkinson DJ, Blinov ML, Begley K, Faeder JR, Gómez HF, Hamm TM, Inagaki Y, Liebermeister W, Lister AL, Lucio D, Mjolsness E, Proctor CJ, Raman K, Rodriguez N, Shaffer CA, Shapiro BE, Stelling J, Swainston N, Tanimura N, Wagner J, Meier-Schellersheim M, Sauro HM, Palsson B, Bolouri H, Kitano H, Funahashi A, Hermjakob H, Doyle JC, Hucka M, Adams RR, Allen NA, Angermann BR, Antoniotti M, Bader GD, Červený J, Courtot M, Cox CD, Dalle Pezze PD, Demir E, Denney WS, Dharuri H, Dorier J, Drasdo D, Ebrahim A, Eichner J, Elf J, Endler L, Evelo CT, Flamm C, Fleming RM, Fröhlich M, Glont M, Gonçalves E, Golebiewski M, Grabski H, Gutteridge A, Hachmeister D, Harris LA, Heavner BD, Henkel R, Hlavacek WS, Hu B, Hyduke DR, Jong H, Juty N, Karp PD, Karr JR, Kell DB, Keller R, Kiselev I, Klamt S, Klipp E, Knüpfer C, Kolpakov F, Krause F, Kutmon M, Laibe C, Lawless C, Li L, Loew LM, Machne R, Matsuoka Y, Mendes P, Mi H, Mittag F, Monteiro PT, Natarajan KN, Nielsen PM, Nguyen T, Palmisano A, Pettit J, Pfau T, Phair RD, Radivojevitch T, Rohwer JM, Ruebenacker OA, Saez-Rodriguez J, Schram M, Schmidt H, Schreiber F, Schubert M, Schulte R, Sealoff SC, Smallbone K, Soliman S, Stefan MI, Sullivan DP, Takahashi K, Teusink B, Tolnay D, Vazirabad I, Kamp A, Wittig U, Wrzodek C, Wrzodek F, Xenarios I, Zhukova A, Zucker J (2020). SBML Level 3: an extensible format for the exchange and reuse of biological models. *Mol Syst Biol* 16(8)

**Kokh DB, Doser B, Richter S, Ormersbach F, Cheng X, Wade RC** (2020). A workflow for exploring ligand dissociation from a macromolecule: Efficient random acceleration molecular dynamics simulation and interaction fingerprint analysis of ligand trajectories. *J. Chem. Phys.* 153(12): p.125-102

**Kozlov A, Alves J, Stamatakis A, Posada D** (2020). *CellPhy*: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data. *bioRxiv* 2020.07.31.230292

**Kramer M, Schneider F, Ohlmann S, Geier S, Schaffenroth V, Pakmor R, Röpke F** (2020). Formation of sdB-stars via common envelope ejection by substellar companions. *A&A* 642:A97

**Lal S, Alpay A, Salzmann P, Cosenza B, Hirsch A, Stawinoga N, Thoman P, Fahringer T, Heuveline V** (2020). SYCL-Bench: A Versatile Cross-Platform Benchmark Suite for Heterogeneous Computing. In: *Euro-Par 2020: Parallel Processing*. Springer International Publishing, pp. 629-644

**Langer N, Schürmann C, Stoll K, Marchant P, Lennon D, Mahy L, Mink Sd, Quast M, Riedel W, Sana H, Schneider P, Schootemeijer A, Wang C, Almeida L, Bestenlehner J, Bodenstein J, Castro N, Clark S, Crowther P, Dufton P, Evans C, Fossati L, Gräfener G, Grassitelli L, Grin N, Hastings B, Herrero A, de Koter Ad, Menon A, Patrick L, Puls J, Renzo M, Sander A, Schneider F, Sen K, Shenar T, Simón-Díaz S, Tauris T, Tramper F, Vink J, Xu X-T** (2020). Properties of OB star-black hole systems derived from detailed binary evolution models. *A&A* 638:A39

**Lutteropp S, Kozlov AM, Stamatakis A** (2020). A fast and memory-efficient implementation of the transfer bootstrap. *Bioinformatics* 36(7): p. 2280-2281

**Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerewinkel N, Mahfouz A, Pinello L, Skums P, Stamatakis A, Attolini CS, Aparicio S, Baaijens J, Balvert M, Barbanson Bd, Cappuccio A, Corleone G, Dutilh BE, Florescu M, Guryev V, Holmer R, Jahn K, Lobo TJ, Keizer EM, Khatri I, Kielbasa SM, Korbel JO, Kozlov AM, Kuo T, Lelieveldt BP, Mandoiu II, Marioni JC, Marschall T, Mölder F, Niknejad A, Raczkowski L, Reinders M, Ridder Jd, Saliba A, Somarakis A, Stegle O, Theis FJ, Yang H, Zelikovsky A, McHardy AC, Raphael BJ, Shah SP, Schönhuth A** (2020). Eleven grand challenges in single-cell data science. *Genome Biology* 21(1) p. 1-35

**López F, Strube M** (2020). A Fully Hyperbolic Neural Model for Hierarchical Multi-class Classification. In Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2020, Online, November 2020, pp. 460-475

**Lösel PD, von de Kamp T, Jayme A, Ershov A, Faragó T, Pichler O, Jerome NT, Aadepu N, Bremer S, Chilingaryan SA, Heethoff M, Kopmann A, Odar J, Schmelze S, Zuber M, Wittbrodt J, Baumbach T, Heuveline V** (2020). Introducing Biomedisa as an open-source online platform for biomedical image segmentation. *Nat Commun* 11(1),5577

**Mahy L, Almeida L, Sana H, Clark J, de Koter Ad, de Mink S, Evans C, Grin N, Langer N, Moffat A, Schneider F, Shenar T, Tramper F** (2020). The Tarantula Massive Binary Monitoring. IV. Double-lined photometric binaries. *A&A* 634:A119

**Mahy L, Sana H, Abdul-Masih M, Almeida L, Langer N, Shenar T, de Koter Ad, de Mink Sd, de Wit S, Grin N, Evans C, Moffat A, Schneider F, Barbá R, Clark J, Crowther P, Gräfener G, Lennon D, Tramper F, Vink J** (2020). The Tarantula Massive Binary Monitoring. III. Atmosphere analysis of double-lined spectroscopic systems. *A&A* 634:A118

**Mathews K, Strube M** (2020). A large harvested corpus of location metonymy. In Proceedings of the 12th International Conference on Language Resources and Evaluation, Marseille, France, May 2020, p.11-16

**Mazzolari A, Nunes-Alves A, Wahab HA, Amaro RE, Cournia Z, Merz KM** (2020). Impact of the Journal of Chemical Information and Modeling Special Issue on Women in Computational Chemistry. *J. Chem. Inf. Model.*, acs.jcim.0c00636

**Miller A. A., Magee M. R., Polin A., Maguire K., Zimmerman E., Yao Y., Sollerman J., Schulze S., Perley D. A., Kromer M., Dhawan S., Bulla M., Andreoni I., Bellm E. C., De K., Dekany R., Delacroix A., Fremling C., Gal-Yam A., Goldstein D. A., Golkhou V. Z., Goobar A., Graham M. J., Irani I., Kasliwal M. M., Kaye S., Kim Y. L., Laher R. R., Mahabal A. A., Masci F. J., Nugent P. E., Ofek E., Phinney E. S., Prentice S. J., Riddle R., Rigault M., Rusholme B., Schweyer T., Shupe D. L., Soumagnac M. T., Terreran G., Walters R., Yan L., Zolkower J., Kulkarni S. R.** (2020). The Spectacular Ultraviolet Flash from the Peculiar Type Ia Supernova 2019yvq, *The Astrophysical Journal*, 898, 56)

**Moreau CA, Quadt KA, Piirainen H, Kumar H, Bhargav SP, Strauss L, Tolia NH, Wade RC, Spatz JP, Kursula I, Frischknecht F** (2020). A function of profilin in force generation during malaria parasite motility independent of actin binding. *J Cell Sci*:jcs.233775

**Morel B, Barbera P, Czech L, Bettsworth B, Hübner L, Lutteropp S, Serdari D, Kostaki E, Mamais I, Kozlov AM, Pavlidis P, Paraskevis D, Stamatakis A** (2020). Phylogenetic analysis of SARS-CoV-2 data is difficult. *Molecular Biology and Evolution*, msaa314

**Mustafa G, Nandekar PP, Mukherjee G, Bruce NJ, Wade RC** (2020). The Effect of Force-Field Parameters on Cytochrome P450-Membrane Interactions: Structure and Dynamics. *Sci Rep* 10(1), pp.72-84

**Müller M** (2020). pyMMAX2: Deep Access to MMAX2 Projects from Python. In Proceedings of the 14th Linguistic Annotation Workshop, Online, December , pp. 167-173.

**Müller M, Ghosh S, Rey M, Wittig U, Müller W, Strube M** (2020). Reconstructing Manual Information Extraction with DB-to-Document Backprojection: Experiments in the Life Science Domain. In Proceedings of the First Workshop on Scholarly Document Processing, Online, November 2020, pp. 81-90. 2020. sdp-1.9

**Nunes-Alves A, Mazzolari A, Merz KM** (2020). What Makes a Paper Be Highly Cited? 60 Years of the Journal of Chemical Information and Modeling. *J. Chem. Inf. Model.* 60(12):5866-5867

**Nunes-Alves A, Kokh DB, Wade RC** (2020). Recent progress in molecular simulation methods for drug binding kinetics. *Current Opinion in Structural Biology* 64:126-133

**Ostaszewski M, Niarakis A, Mazein A, Kuperstein I, Phair R, Orta-Resendiz A, Singh V, Aghamiri SS, Acencio ML, Glaab E, Ruepp A, Fobo G, Montrone C, Brauner B, Frischman G, Gómez LCM, Somers J, Hoch M, Gupta SK, Scheel J, Borlinghaus H, Czauderna T, Schreiber F, Montagud A, Leon MPd, Funahashi A, Hiki Y, Hiroi N, Yamada TG, Dräger A, Renz A, Naveez M, Bocskei Z, Messina F, Börnigen D, Fergusson L, Conti M, Rameil M, Nakonecní V, Vanhoefer J, Schmiester L, Wang M, Ackerman EE, Shoemaker J, Zucker J, Oxford K, Teuton J, Kocakaya E, Summak GY, Hanspers K, Kutmon M, Coort S, Eijssen L, Ehrhart F, Rex DAB, Slenter D, Martens M, Haw R, Jassal B, Matthews L, Orlic-Milacic M, Ribeiro AS, Rothfels K, Shamovsky V, Stephan R, Sevilla C, Varusai T, Ravel J, Fraser R, Ortseifen V, Marchesi S, Gawron P, Smula E, Heirendt L, Satagopam V, Wu G, Riutta A, Golebiewski M, Owen S, Goble C, Hu X, Overall RW, Maier D, Bauch A, Gyori BM, Bachman JA, Vega C, Grouès V, Vazquez M, Porras P, Licata L, Iannuccelli M, Sacco F, Nesterova A, Yuryev A, Waard Ad, Turei D, Luna A, Babur O, Soliman S, Valdeolivas A, Esteban-Medina M, Peña-Chilete M, Helikar T, Puniya BL, Modos D, Treveil A, Olbei M, Meulder BD,**

**Dugourd A, Naldi A, Noel V, Calzone L, Sander C, Demir E, Korcsmaros T, Freeman TC, Augé F, Beckmann JS, Hasenauer J, Wolkenhauer O, Wilighagen EL, Pico AR, Evelo CT, Gillespie ME, Stein LD, Hermjakob H, D'Eustachio P, Saez-Rodriguez J, Dopazo J, Valencia A, Kitano H, Barillot E, Auffray C, Balling R, Schneider R** (2020). COVID-19 Disease Map, a computational knowledge repository of SARS-CoV-2 virus-host interaction mechanisms. *bioRxiv* 2020.10.26.356014

**Öztürk MA, Wade RC** (2020). Computation of FRAP recovery times for linker histone – chromatin binding on the basis of Brownian dynamics simulations. *Biochimica et Biophysica Acta (BBA) - General Subjects* 1864(10):129653

**Öztürk MA, De M, Cojocaru V, Wade RC** (2020). Chromosome Structure and Dynamics from Molecular Simulations. *Annu. Rev. Phys. Chem.* 71(1):101-119

**Rennekamp B, Kutzki F, Obarska-Kosinska A, Zapp C, Gräter F** (2020). Hybrid Kinetic Monte Carlo/Molecular Dynamics Simulations of Bond Scissions in Proteins. *J. Chem. Theory Comput.* 16(1):553-563

**Sadiq SK** (2020). Fine-Tuning of Sequence Specificity by Near Attack Conformations in Enzyme-Catalyzed Peptide Hydrolysis. *Catalysts* 10(6):684

**Sand C, Ohlmann ST, Schneider FRN, Pakmor R, Röpke FK** (2020). Common-envelope evolution with an asymptotic giant branch star. *A&A* 644:A60

**Schneider FRN, Ohlmann ST, Podsiadlowski P, Röpke FK, Balbus SA, Pakmor R** (2020). Long-term evolution of a magnetic massive merger product. *Monthly Notices of the Royal Astronomical Society* 495(3):2796-2812

**Schreiber F, Sommer B, Czauderna T, Golebiewski M, Gorochowski TE, Hucka M, Keating SM, König M, Myers C, Nickerson D, Waltemath D** (2020). Specifications of standards in systems and synthetic biology: status and developments in 2020. *Journal of Integrative Bioinformatics* 17(2-3)

**Shcherbakov O, Polsterer K, Svyatny VA** (2020). Integration of Distributed Parallel Simulation Environment with Cloud-Infrastructures, *PMDA* 22(1):22-27

**Suyu SH., Huber S., Cañameras R, Kromer M, Schultdt S, Taubenberger S, Yıldırım A, Bonvin V, Chan JHH, Courbin F, Nöbauer U, Sim SA, Sluse D** (2020). HOLISMOKEs. I. Highly Optimised Lensing Investigations of Supernovae, Microlensing Objects, and Kinematics of Ellipticals and Spirals, *Astronomy & Astrophysics*, 655, A162

# 8 Teaching

**van Son LAC, De Mink SE, Broekgaarden FS, Renzo M, Justham S, Laplace E, Morán-Fraile J, Hendriks DD, Farmer R** (2020). *Polluting the Pair-instability Mass Gap for Binary Black Holes through Super-Eddington Accretion in Isolated Binaries*. ApJ 897(1):100

**Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., & Gneiting, T.** (2020). *Skill of Global Raw and Postprocessed Ensemble Predictions of Rainfall in the Tropics*, Weather and Forecasting, 35(6), 2367-2385

**Waltemath D, Golebiewski M, Blinov ML, Gleeson P, Hermjakob H, Hucka M, Inau ET, Keating SM, König M, Krebs O, Malik-Sheriff RS, Nickerson D, Oberortner E, Sauro HM, Schreiber F, Smith L, Stefan MI, Wittig U, Myers CJ** (2020). *The first 10 years of the international coordination network for standards in systems and synthetic biology (COMBINE)*. Journal of Integrative Bioinformatics 17(2-3)

**Weidner P, Söhn M, Schroeder T, Helm L, Hauber V, Gutting T, Betge J, Röcken C, Rohrbacher FN, Pattabiraman VR, Bode JW, Seger R, Saar D, Nunes-Alves A, Wade RC, Ebert MPA, Burgermeister E** (2020). *Myotubularin-related protein 7 activates peroxisome proliferator-activated receptor-gamma*. Oncogenesis 9(6), 59

**Wu S, Everson RW, Schneider FRN, Podsiadlowski P, Ramirez-Ruiz E** (2020). *The Art of Modeling Stellar Mergers and the Case of the B[e] Supergiant R4 in the Small Magellanic Cloud*. ApJ 901(1):44

**Yuan J-H, Han SB, Richter S, Wade RC, Koch DB** (2020). *Druggability Assessment in TRAPP Using Machine Learning Approaches*. J. Chem. Inf. Model. 60(3):1685-1699

**Zapletal A, Höller D, Sinz C, Stamatakis A** (2020). *SoftWipe – a tool and benchmark to assess scientific software quality*. bioRxiv 2020.10.07.330621

**Zapp C, Obarska-Kosinska A, Rennekamp B, Kurth M, Hudson DM, Mercadante D, Barayeu U, Dick TP, Denysenkov V, Prisner T, Bennati M, Daday C, Kappl R, Gräter F** (2020). *Mechanoradicals in tensed tendon collagen as a source of oxidative stress*. Nat Commun 11(1), 2315

**Zhu J-Y, Song C, Heuveline V, Li B, Li B-H, Han Z-S, Liu X-Y** (2020). *Characterization Simulation of a Bulk MOSFET in Steady-State with SIPG Method*. IEEE 15th International Conference on Solid-State & Integrated Circuit Technology (ICSICT), pp.1-3, IEEE

## Degrees

### Jonas Brehmer:

"*Theory and Methodology of Scoring Functions: Tail Properties, Interval Forecasts, and Point Processes*", Ph.D. thesis, Faculty of Business Informatics and Business Mathematics, University of Mannheim (Martin Schlather, Kirstin Strokorb), and HITS: Tilmann Gneiting (2020).

### David Bubeck:

"*Thermonuclear Explosions of Rapidly Rigidly Rotating White Dwarfs*", Master's thesis, Department of Physics and Astronomy, Heidelberg University, and HITS: Friedrich Röpke (2020).

### Thomas Ehret:

"*Comparative Study of FERM-PI(4,5)P2 Interaction Dynamics of Ezrin and Focal Adhesion Kinase*", Master's thesis, Heidelberg University, Faculty of Physics, Frauke Gräter and Ulrich Schwarz (2020).

### Philipp Gerstner:

"*Analysis and Numerical Approximation of Dielectrophoretic Force Driven Flow Problems*." Ph.D. thesis, Faculty of Mathematics and Computer Science, Heidelberg University, and HITS: Vincent Heuveline (2020).

### Dimitri Höller:

"*Advanced Heuristics for Accelerating PaPaRa*", Master's thesis, Karlsruhe Institute of Technology, and HITS: Alexandros Stamatakis (2020).

### Johannes Horn:

"*Singular fibers of Hitchin systems*", Ph.D. Thesis, Mathematics Institute, Heidelberg University, and HITS: Daniele Alessandrini, Anna Wienhard (2020).

### Lukas Hübner:

"*Load-Balance and Fault-Tolerance for Massively Parallel Phylogenetic Inference*", Master's thesis, Karlsruhe Institute of Technology, and HITS: Alexandros Stamatakis (2020).

### Lukas Jarosch:

"*Computational modeling of SERCA interactions with S100A1ct and DWORF fingerprints*", Bachelor's thesis, Biochemistry, Faculty of Biosciences and Faculty of Chemistry and Geosciences, Heidelberg University, and HITS: Manuel Glaser and Rebecca C. Wade (2020).

### Mohsen Mesgar:

"*Graph-based Patterns for Local Coherence Modeling*", Ph.D. Thesis, Neuphilologische Fakultät, Heidelberg University, and HITS: Michael Strube (2020).

### Fabian Ormersbach:

"*Computational ligand superimposition and analysis of protein-ligand interaction fingerprints*", Bachelor's thesis, Molecular Biotechnology, Faculty of Biosciences, Heidelberg University, and HITS: Daria Kokh and Rebecca C. Wade (2020).

### Ina A. Pöhner:

"*Computational approaches to drug design against the folate & biopterin pathways of parasites causing neglected tropical diseases*", Ph.D. thesis, Combined Faculty for the Natural Sciences and Mathematics, Heidelberg University, and HITS: Rebecca C. Wade (2020).

### Lucas Rettenmeier:

"*Word Embeddings: Stability and Semantic Change*", Master's thesis, Department for Physics and Astronomy, Heidelberg University (Fred Hamprecht), and HITS: Michael Strube (2020).

### Eugen Rogozinnikov:

"*Symplectic groups over noncommutative rings and maximal representations*", Ph.D. Thesis, Mathematics Institute, Heidelberg University, and HITS: Daniele Alessandrini, Anna Wienhard (2020).

### Paul Schade:

"*Phylogenetic species tree inference from gene trees despite paralogy*", Master's thesis, Karlsruhe Institute of Technology, and HITS: Alexandros Stamatakis (2020).

### Anna Schroeder:

"*Collagen Structure and Mechanics: Molecular Dynamics Simulations and Network Analysis*", Master's thesis, Heidelberg University, Faculty of Physics, Frauke Gräter and Ulrich Schwarz (2020).

## Lectures, Courses and Seminars

### Nguyen-Thi Dang:

"*The top Lyapunov exponent*", Heidelberg University, summer semester 2020.

### Timo Dimitriadis (with Robert Jung):

Lecture and exercise course on "Applied Financial Econometrics", Universität Hohenheim, summer semester 2020.

### Timo Dimitriadis:

Lecture and exercise course on "Allgemeine Methodenlehre der Statistik (Statistical methods)", Heidelberg University, winter semester 2020/2021.

### Dorotea Dudas, Xiaoming Hu, Maja Rey, Andreas Weidemann and Ulrike Wittig:

de.NBI Course "Tools for Systems biology modeling and data exchange: COPASI, CellNetAnalyzer, SABIO-RK, FAIRDOMHub/SEEK", online, 21 - 23 September 2020.

### Javier Moran Fraile:

Tutorial course "Computational Astrophysics", Heidelberg University, summer semester 2020.

### Tilmann Gneiting (lectures) and Johannes Resin (exercises):

Course on "Time Series Analysis", Karlsruhe Institute of Technology, summer semester 2020.

Course on "Forecasting: Theory and Practice I", Karlsruhe Institute of Technology, winter semester 2020/2021.

### Martin Golebiewski and Wolfgang Müller:

"*LiSyM Data Management*", LiSyM Retreat 2020 of the Liver Systems Medicine Network, Hofgeismar, Germany, 29-31 January 2020.

### Frauke Gräter:

Contribution to lecture "Computational biochemistry" for biochemistry master students, winter semester 2020/21.

Contribution to lecture "Biophysical Chemistry" for Molecular Biotechnology bachelor students, winter semester 2020/21. Lecture with practicals "Physical Chemistry of Life II" winter semester 2020/21 (with Michael Boutros).

**Frauke Gräter, Fabian Kutzki and Benedikt Rennekamp:**  
Lecture with tutorials "Fundamentals of simulation methods", winter semester 2020/21 (with Friedrich Röpke).

**Frauke Gräter and Camilo Aponte-Santamaría:**  
"Data Science 2020", Max Planck School Matter to Life, Heidelberg, Germany, October-December 2020.

**Frauke Gräter (MBM), Rebecca C. Wade (MCM):**  
M.Sc. Seminar course "Machine Learning for the Biomolecular World", Heidelberg University, summer semester, 2020.

**Frauke Gräter (MBM), Rebecca C. Wade, Ariane Nunes-Alves, Manuel Glaser (MCM):**  
M.Sc. practical course "Computational Molecular Biophysics", Heidelberg University, winter semester, 2019/2020.

**Ganna Gryn'ova and Christopher Ehler:**  
Special Lecture Course "Applied Computational Chemistry", Heidelberg University, summer semester 2020.

**Olga Krebs:**  
"FAIR Principles for research data management with classroom discussions", Lecture and live demo at the ELIXIR Luxembourg's training on "Best practices in research data management and stewardship", 3 June 2020. "FAIR data principles and Data publishing and archival", Lecture and practical session at the ELIXIR Luxembourg+ELIXIR France online training , 5 - 8 October 2020. "FAIR data training" computer practicals at the Modelling COVID-19 epidemics course , 30 November - 9 December 2020.

**Giovanni Leidi:**  
Tutorial accompanying the lecture course "Computational Astrophysics", Heidelberg University, summer semester 2020. Tutorial accompanying the lecture course "Fundamentals of Simulation Methods", Heidelberg University, winter semester 2020/2021.

**Sebastian Lerch:**  
Lecture course on "Probability and Statistics for Mechanical Engineering", Karlsruhe Institute of Technology, summer semester 2020. Lecture course on "Probability and Statistics for Computer Scientists", Karlsruhe Institute of Technology, winter semester 2020/2021.

**Sebastian Lerch (with Joaquim Pinto):**  
Lecture course on "Methods of Data Analysis", Karlsruhe Institute of Technology, summer semester 2020.

**Sebastian Lerch and Eva-Maria Walz:**  
Short course "Introduction to Machine Learning", Karlsruhe Institute of Technology, Graduate School of Center MathSEE, summer semester 2020.

**Mark-Christoph Müller:**  
Seminar: "Multiword Expressions", Department of Computational Linguistics, Heidelberg University (Winter Semester 2019/2020).

**Maria Beatrice Pozzetti:**  
"Translation Surfaces", Heidelberg University, Heidelberg, summer semester 2020; "Differential Geometry 2", Heidelberg University, winter semester 2020/2021; "Bridgeland's stability for meromorphic differentials", Heidelberg University, winter semester 2020/2021.

**Maja Rey and Ulrike Wittig:**  
FAIRDOM-SEEK Training at MIX-UP Kick off Meeting, Brussels, Belgium, 7 February 2020.

**Friedrich Röpke:**  
Lecture course "Computational Astrophysics", Heidelberg University, summer semester 2020. Lecture course "The Stellar Cookbook: A practical guide to the theory of stars", Heidelberg University, winter semester 2019/2020 and winter semester 2020/2021, together with Fabian Schneider. Lecture course "Fundamentals of Simulation Methods", Heidelberg University, winter semester 2020/2021, together with Frauke Gräter. Research Seminar "Physics of Stellar Objects", Heidelberg University, winter semester 2019/2020, summer semester 2020, and winter semester 2020/2021.

**Christian Sand:**  
Exercises and tutorial accompanying the lecture course "Physik A", Heidelberg University, winter semester 2019/2020.

**Anna-Sofie Schilling:**  
"Fun fact aus Analysis und Linearer Algebra", Heidelberg University, winter semester 2020/2021.

**Fabian Schneider:**  
Lecture course "The Stellar Cookbook: A practical guide to the theory of stars", Heidelberg University, winter semester 2020/2021, together with Friedrich Röpke. Lecture

course "The Stellar Cookbook: A practical guide to the theory of stars", Heidelberg University, winter semester 2019/2020, together with Friedrich Röpke. Master Seminar "Cosmic explosions", Heidelberg University, summer semester 2020. Lecture course "Python: programming for scientists", Heidelberg University, summer semester 2020. Lecture course "Python: programming for scientists", Heidelberg University, winter semester 2019/2020.

**Theodoros Soulannis:**  
Tutorial course "Experimental Physics II", Heidelberg University, summer semester 2020.

**Alexandros Stamatakis, Ben Bettisworth, Alexey Kozlov, Pierre Barbera:**  
Lecture "Introduction to Bioinformatics for Computer Scientists", computer science Master's program at Karlsruhe Institute of Technology, winter semester, 2019/2020.

**Alexandros Stamatakis, Benoit Morel, Alexey Kozlov, Pierre Barbera:**  
Lecture "Introduction to Bioinformatics for Computer Scientists", computer science Master's program at Karlsruhe Institute of Technology, winter semester, 2020/2021.

**Alexandros Stamatakis, Benoit Morel, Pierre Barbera, Ben Bettisworth, Alexey Kozlov:**  
Seminar "Hot Topics in Bioinformatics", computer science Master's program at Karlsruhe Institute of Technology, summer semester, 2020.

**Alexandros Stamatakis:**  
EU ITN IGNITE project PhD student training lectures on "General Introduction to Markov Chains", "Green high performance computing", "Ascertainment bias correction in phylogenetics", "Discrete Operations on phylogenetic trees", "Introduction to parsimony and its computational commonalities with likelihood", "High Performance Computing and Phylogenetics", "Introduction to Bayesian phylogenetics using MCMC", "Computing the Phylogenetic Likelihood", on-line, August 2020.

**Michael Strube:**  
PhD Colloquium, Department of Computational Linguistics, Heidelberg University (Winter Semester 2019/2020).  
PhD Colloquium, Department of Computational Linguistics, Heidelberg University (Summer Semester 2020).

### **Gabriele Viaggi:**

"Hyperbolic manifolds", Heidelberg University, winter semester 2020/2021.

**Rebecca C. Wade, Christina Athanasiou, Manuel Glaser, Daria Kokh, Abraham Muniz, Ariane Nunes-Alves, Stefan Richter, Athanasios-Alexandros Tsengenes, (MCM), Frauke Gräter, Isabel Martin, Svenja de Buhr (MBM):**

B.Sc. Biosciences practical course "Grundkurs Bioinformatik", Heidelberg University, 20-24 January 2020.

**Rebecca C. Wade, Christina Athanasiou, Athanasios-Alexandros Tsengenes:**

EuroNeurotrophin International Training Network (ITM) 3rd Training Week "Structural Biology Approaches for Neurodegeneration", held online, 30 September – 1 October 2020.

### **Rebecca C. Wade:**

Module 4, "Biomolecular Recognition: Modeling and Simulation", M.Sc. Molecular Cell Biology, Heidelberg University, 13 March 2020. Module 3, "Protein Modeling", M.Sc. Molecular Cell Biology, Heidelberg University, 20 & 25 May 2020. Ringvorlesung "Computational Biochemistry", "Electrostatics and Solvation for Biomolecules", M.Sc. Biochemistry, Heidelberg University, 30 October 2020. Ringvorlesung „Biophysik“, "Receptor-Ligand Interactions: Structure and Dynamics", B.Sc. Molecular Biotechnology, Heidelberg University, 15 December 2020. Ringvorlesung „Mobi4all“, "Computational Approaches to Protein Dynamics and Binding Kinetics for Drug Discovery", M.Sc. Molecular Biotechnology, Heidelberg University, December, 2020.

### **Anna Wienhard:**

"Topological methods in Data Analysis", Journal Club, Heidelberg University, summer semester 2020; "Lineare Algebra I", Heidelberg University, winter semester 2020/2021.

# 9 Miscellaneous

## 9.1 Guest Speaker Activities (invited only)

### Johannes Bracher:

"Assembling, Comparing and Combining COVID-19 forecasts", Frankfurt Institute for Advanced Studies (FIAS) M<sup>2</sup>C Seminars (online), Frankfurt, Germany, 28 October 2020.

### Nguyen-Thi Dang:

"Mélange topologique des flots hyperboliques homogène", HORUS seminar, IRMA, Strasbourg, France, 5 June 2020; "Topological mixing of the Weyl chamber", Ergodic theory and dynamical systems Seminar, Bristol, UK, 29 October 2020.

### Madhura De:

"Single-molecule studies on mono- and trinucleosomes", Heidelberg Chromatin Club, DKFZ, Heidelberg, 27 February 2020. "Single-molecule studies on mono- and trinucleosomes" Institute of Biological and Chemical Systems, Karlsruhe Institute of Technology, Karlsruhe, Germany, 2 April 2020.

### Antonio D'Isanto:

"ESCAPE – Building the infrastructure for the next generation astronomy", Invited seminar for the Astrophysics group journal club, University of Naples Federico II, Physics Department, Naples, Italy, 7 January 2020. "Convolutional neural networks: Theory and application in astronomy", Invited lecture for the Astroinformatics course, University of Naples Federico II, Physics Department, Naples, Italy, 8 January 2020. "The two worlds of photometric redshift estimation via machine learning: fully automatic vs feature based", Invited seminar for the Machine learning in Astronomy meetings of the West Sydney University, Australia (online), 9 June 2020; invited lecture for the Machine learning for astronomers course at the Masaryk's University, Brno, Czech republic (online), 24 November 2020. "MEGAVIS - Real-time spectra analysis and visualization with autoencoders", Invited talk for the PUNCHLunch Webinar series (online), Leibniz Institute for Astrophysics Potsdam (AIP), Germany, 3 December 2020.

### Valentina Disarlo:

"Lignes d'étiirement de Thurston généralisées", Séminaire GT3, Université de Strasbourg, Strasbourg, France, 20 January 2020; "Lignes d'étiirement de Thurston généralisées", Séminaire Géométrie, Université Sorbonne (Jussieu), Paris, 20 February 2020.

### Tilmann Gneiting:

"Isotonic Distributional Regression – Leveraging Monotonicity, Uniquely So!", Computational Science Lab (CSL) Symposium, Universität Hohenheim, Hohenheim, Germany, 22 January 2020. "Wettervorhersage: Welche Rolle spielt der Zufall?", Carl-Bosch-Museum, Heidelberg, Germany, 29 January 2020. "Isotonic Distributional Regression (IDR) – Leveraging Monotonicity, Uniquely So!", Luminy Conference Mathematical Methods of Modern Statistics 2, Luminy, France (online), 16 June 2020. "Erfolge und Grenzen der Wettervorhersage – eine mathematische Perspektive", Universität Ulm, Ulm, Germany (online), 10 July 2020.

### Martin Golebiewski:

ELIXIR Guest Webinar: "Two universes, one world - Community standards vs. formal ISO standards in the life sciences", ELIXIR webinar series (online), 16 January 2020. "Establishing and harmonizing technical standards and key performance indicators for human digital twins", European Commission Workshop 'Human Digital Twin' (online), 6 November 2020.

### Ganna Gryn'ova:

"Happy Together: How Computational Chemistry Can Advance Functional Materials", seminar talk at Centre for Advanced Materials, Heidelberg University, Heidelberg, Germany, 12 February 2020; "Crossing Electronic Bridges: Computational Chemistry of Molecular Junctions", (Bio) Molecular Electronics Colloquia (online), University of Liverpool, UK, 1 October 2020.

### Johannes Horn:

"Singular fibers of Hitchin integrable systems", Seminar Symplectic Geometry, Gauge Theory and Categorification, Columbia University, New York, USA, 21 February 2020; "Singular fibers of Hitchin integrable systems", Geometry-Topology Seminar, University of Maryland, College Park, USA, March 2020.

### Daria Kokh:

"Exploring ligand unbinding kinetics using random acceleration molecular dynamics: what can we learn?", "2020 Workshop on Free Energy Methods in Drug Design", 12-14 November 2020

### Alexey Kozlov:

"New phylogenetic tools for genome-scale datasets: RAxML-NG, ParGenes and GeneRax", 2nd Bird 10,000 Genomes (B10K) project symposium and Workshop, Copenhagen, Denmark, February 2020. "Random thoughts on GreenHPC", Sustainability meeting at the Max Planck Institute for Astronomy", Heidelberg, Germany (online), December 2020.

### Olga Krebs:

"FAIRDOMHub: Implementing FAIR Data Principles for scientific data management and stewardship", Lecture at the Modelling COVID-19 epidemics course, 1 December 2020.

### Sebastian Lerch:

"Neural networks for postprocessing ensemble weather forecasts", National Oceanic and Atmospheric Administration (NOAA) Workshop on Leveraging AI in Environmental Sciences, College Park, USA (online), 22 October 2020. "Predictive Inference Based on Markov Chain Monte Carlo Output", Conference on Information Technology and Data Science, Debrecen, Hungary (online), 7 November 2020.

### Brice Loustau:

"The hyper-Kähler geometry of minimal hyperbolic germs", Geometry seminar, University of Wisconsin at Milwaukee, USA, 9 November 2020.

### Benoit Morel:

"Gene tree inference under gene duplication, transfer and loss with GeneRax", Sanger Institute, Cambridge, UK, Microbial Genomics Meetings (online), December 2020.

### Goutam Mukherjee:

"RASPD+: Fast protein-ligand binding free energy prediction using machine learning and its applications to SARS-CoV-2 targets", Drug Discovery Hackathon 2020 workshop, <https://www.youtube.com/watch?v=9WarUhvNpGg&t=162s>, 18 August 2020. "RASPD+: Machine Learning for Fast Protein-Ligand Binding Free Energy Prediction and its Applications to SARS-CoV-2 Targets", BMS Institute of Technology and Management Bengaluru, India (International Webinar), 3 October 2020.

### Wolfgang Müller:

Topic Lead Excel Validator, Integrative Pathway Modelling in Systems Biology and Systems Medicine Hackathon, Bonn, Germany, 5-7 February 2020. "FAIR data in de.NBI" de.NBI plenary, Berlin, Germany, 14 February 2020.

### Mareike Pfeil:

"Deforming Anosov representations: From earthquakes maps to cataclysms", Geometry Graduate Colloquium, Zürich, Switzerland, 2 April 2020.

### Maria Beatrice Pozzetti:

"Real spectrum compactification of character varieties", Strasbourg, 15 June 2020; "What's new in Higher (rank) Teichmüller theory?", ETH Zürich, Switzerland, 30 October 2020; "Orbit growth rate and Hausdorff dimensions for Anosov representations", Paderborn, Germany, 3 November 2020; "The real spectrum compactification of higher rank Teichmüller spaces", Frankfurt, Germany, 18 November 2020.

### Anja Randecker:

"Random walks on hyperbolic manifolds", Oberseminar Algebra und Zahlentheorie, Saarbrücken, Germany, 3 February 2020; "Large-scale geometry of the saddle connection complex", Séminaire Teich, Marseille, France (online), December 2020.

### Friedrich Röpke:

"A 3D view on stellar astrophysics", Heidelberg Joint Astronomical Colloquium, Heidelberg, Germany, 21 January 2020. "Type Ia supernova explosion models: challenges and implications", Meeting of the Royal Astronomical Society, London, UK (online), 13 November 2020.

### Carmen Rovi:

"Whitney stratified spaces", Heidelberg Topology seminar, Heidelberg University, Germany, 30 January 2020; "Topology meets Physics: Scissors congruences and TQFTs", Colloquium at Harvey Mudd College, Claremont, California, USA, 13 February 2020; "The Whitehead group of a CW complex", Heidelberg Topology seminar, Heidelberg University, Germany, 7 May 2020; "The torsion of a homotopy equivalence", Heidelberg Topology seminar, Heidelberg University, Germany, 23 July 2020; "Introduction to Surgery Theory", RTG lectures, Heidelberg, Germany, (online), November-December 2020; "Kontsevich's configuration space integrals and characteristic classes of framed sphere bundles", Topology seminar, Münster, Germany, 18 December 2020.

### Fabian Schneider:

"State-of-the-art evolution models for single, binary and magnetic massive stars", Invited talk at Conference "MOBSTER-I: Stellar variability as a probe of magnetic fields in massive stars" (online), 16 July 2020. "The strongest magnets in the Universe", ARI Colloquium, Astronomisches Rechen-Institut, Centre for Astronomy, Heidelberg University, Heidelberg, Germany (online), 26 November 2020.

**Alexandros Stamatakis, Alexey Kozlov:**

"RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference", ISCB International Society for Computational Biology, webinar (online), September 2020.

**Rebecca C. Wade:**

"Tau-RAMD tutorial: Fast estimation of drug residence times", 2020 MolSSI School on "Open Source Software for Rare Event Sampling Strategies", <https://github.com/westpa/westpa/wiki/2020-MolSSI-School-on-Open-Source-Software-in-Rare-Event-Path-Sampling-Strategies>, 15 July 2020.

"Zooming in on the dynamic interactions of proteins and drugs by computer simulation", CRC 1093 Lecture Series, University of Duisburg-Essen, Germany (online), 17 November 2020. "Computational Approaches to Protein Dynamics and Binding Kinetics for Drug Discovery", Molecular Graphics and Modelling Society (MGMS) Lecture Tour Lecture, <https://www.mgms.org/WordPress/lecture-tour/>, <https://www.youtube.com/watch?v=QFbxeovFRtg&t=101s> (online), 24 November 2020. "Zooming in on the dynamic interactions of proteins and drugs by computer simulation", "Science at the Edge" colloquium, Biochemistry, Physics and Chemical Engineering / Materials Sciences Departments, Michigan State University, USA (online), 4 December 2020.

**Anna Wienhard:**

"Geometry, Topology, and Groups", Mathematics Colloquium, IST Austria, 11 March 2020; "Where geometry meets dynamics: groups, entropy and Hausdorff dimension", Mathematical Colloquium, Rutgers University, USA, 30 September 2020; "Where geometry meets dynamics: groups, entropy and Hausdorff dimension", Mathematics Colloquium, Florida State University, USA, 13 November 2020; "Growth of groups, entropy and Hausdorff dimension", Zurich Colloquium in Mathematics, Switzerland, 24 November 2020.

**Ulrike Wittig:**

"Research data management by SEEK/FAIRDOMHub", Seminar "Special Interest Group Data Infrastructure", Stuttgart, Germany, 5 February 2020. "SABIO-RK - a manually curated reaction kinetics database", South Africa-Germany bilateral research projects workshop on "Systematic profiling of multicopper oxidases" (online), 5-7 October 2020.

**9.2 Presentations (contributed talks and posters)****Robert Andrassy:**

"Fully compressible simulations of wave generation and propagation in main-sequence stars", talk at the Annual Meeting of the European Astronomical Society, Leiden, Netherlands (online), 1 July 2020. "Convective boundary mixing in stellar interiors", talk at the 15th Würzburg Workshop, Heidelberg, Germany (online), 14 December 2020.

**Pierre Barbera:**

"Phylogenetic Placement: Why the where of the what tells us the who and the why", University of Duisburg-Essen, Essen, Germany, February 2020.

**Thomas Baumann:**

"Common Envelope Parameter Space Exploration", talk at the 15th Würzburg Workshop, Heidelberg, Germany (online), 15 December 2020.

**Johannes Bracher:**

"Assembling, Comparing and Combining COVID-19 forecasts", Karlsruhe Institute of Technology Center for Information, Systems, Technology (KCIST), Germany, Online Lecture Series, 26 October 2020.

**Jonas Brehmer:**

"Scoring Interval Forecasts", Bernoulli-IMS One World Symposium 2020 (online), 24-28 August 2020.

**Nguyen-Thi Dang:**

"Topological dynamics of the Weyl chamber flow", Sophus Lie 2020, Paderborn, Germany, February 2020; "Topological dynamics of the Weyl chamber flow", Nearly Carbon Neutral Geometric Topology Conference (online), June 2020.

**Timo Dimitriadis:**

"Testing Forecast Rationality for Measures of Central Tendency", Computational Science Lab (CSL) Symposium, Universität Hohenheim, Hohenheim, Germany, 22 January 2020; Alfred-Weber-Institut, Universität Heidelberg, Heidelberg, Germany, 23 January 2020; 12th Econometric Society World Congress, Bocconi University, Milan, Italy (online), 19 August 2020. "Evaluating Probabilistic Classifiers: Reliability Diagrams and Score Decompositions Revisited", International Symposium on Forecasting, Rio de Janeiro, Brazil (online), 28 October 2020; Alfred-Weber-Institut, Universität Heidelberg, Heidelberg, Germany (online), 18 November 2020.

**Antonio D'Isanto:**

"Real-time spectra analysis and visualization with autoencoders", ESCAPE WP4 Tech Forum 1, Strasbourg, France, 4 – 6 February 2020; IVOA Virtual Interop 2020, Virtual conference, 4 – 8 May 2020; AG 2020 – Meeting of the German Astronomical Society, Virtual conference, 21 – 25 September 2020.

**Dorotea Dudas:**

"SABIO-VIS", lightning talk at HARMONY 2020, EMBL-EBI, Hinxton, UK, 9-13 March 2020.

**Christopher Ehrlert:**

"An Assessment of Cluster Models and Methods", Graphene2020 (online), 19 - 23 October 2020 (poster).

**Florian Franz and Frauke Gräter:**

"Talin impacts force-induces Vinculin activation through "loosening" the vinculin inactive state", Annual Meeting of the Biophysical Society, San Diego, USA, 15-19 February 2020 (poster).

**Javier Moran Fraile:**

"3D MHD simulations of White Dwarf - Neutron Star mergers", talk at the 15th Würzburg Workshop, Heidelberg, Germany (online), 15 December 2020.

**Martin Golebiewski:**

"Options for standardization strategies in personalized medicine", Workshop "Using patient derived data for in silico modelling in personalized medicine", HITS, Heidelberg, Germany, 4 February 2020. "FAIR Data Infrastructures for Biomedical Research", Annual Conference of the Research Data Alliance Germany (RDA Deutschland), Potsdam, Germany, 25-27 February 2020. "A European standardization framework for data integration and data-driven in silico models for personalized medicine", lightning talk at HARMONY 2020, EMBL-EBI, Hinxton, UK, 9-13 March 2020. "Chances and challenges in developing standards", EU-STANDS4PM annual meeting 2020 (online), 27 May 2020. "ISO and community standards for Modelling in personalized Medicine", COMBINE 2020, session "Towards in silico approaches for personalized medicine" (online), 7 October 2020.

"FAIRness of data and models through standardization in systems biology", COMBINE 2020, session "Data, tools, standards and more: infrastructure for systems biology" (online), 8 October 2020. "FAIR Data Infrastructures for Biomedical Informatics", introduction for GMDS workshop, 65th Annual Conference of the German Association for Medical Informatics, Biometry and Epidemiology (GMDS) (online), 15 October 2020. "Standards for FAIR Data", NFDI4Health Kick-off-Meeting (online), 4 November 2020. "SEEK and Find: FAIRDOM Data Management Support for COVID-19 Disease Maps", 5th Disease Maps Community Meeting (online), 12-14 November 2020.

**Sabrina Gronow:**

"Double detonations", talk at the 15th Würzburg Workshop, Heidelberg, Germany (online), 16 December 2020.

**Johann Higl:**

"Calibrating the Core Overshooting Parameter With Hydrodynamical Simulations", poster at the Annual Meeting of the European Astronomical Society, Leiden, Netherlands (online), 1 July 2020. "Low Mach Number Simulations and Their Need for Well-Balancing", talk at the Annual Meeting of the German Astronomical Society (online), 25 September 2020. "Compressible Simulations of Stellar Oscillations", talk at the 15th Würzburg Workshop, Heidelberg, Germany (online), 14 December 2020.

**Xiaoming Hu:**

"SEEK : A systems biology data and model management platform", lightning talk at HARMONY 2020, EMBL-EBI, Hinxton, UK, 9-13 March 2020.

**Alexander Jordan:**

"Evaluating Probabilistic Classifiers: Reliability Diagrams and Score Decompositions Revisited", International Verification Methods Workshop Online, 18 November 2020.

**Markus Kurth and Frauke Gräter:**

"A role for dihydroxyphenylalanine (DOPA) and its radical as marker for mechanical stress in collagen", EMBO Workshop: Chemical Biology 2020 (online), Heidelberg, Germany, 3-5 September 2020 (poster).

**Florian Lach:**

"Nucleosynthesis Imprints of different Type Ia Supernova Explosion Scenarios and Chandrasekhar-mass Deflagrations in Carbon-Oxygen White Dwarfs", talk at the 15th Würzburg Workshop, Heidelberg, Germany (online), 16 December 2020.

**Giovanni Leidi:**

"Implementing MHD into SLH", talk at the 15th Würzburg Workshop, Heidelberg, Germany (online), 16 December 2020.

**Sebastian Lerch:**

"Neural Networks for Postprocessing Ensemble Weather Forecasts", EUMETNET Workshop on AI for Weather and Climate Modeling, Brussels, Belgium, 2 February 2020; ECMWF-ESA Workshop on Machine Learning for Earth System Observation and Prediction, Reading, UK (online), 7 October 2020. "Evaluating Probabilistic Forecasts with scoringRules", International Verification Methods Workshop Online, 18 November 2020. "Artificial Intelligence for Probabilistic Weather Prediction", Karlsruhe Institute of Technology Center for Information, Systems, Technology (KCIST) Online Lecture Series, Germany, 7 December 2020.

**Kiril Maltsev:**

"On the foundation of Black Hole Thermodynamics", talk at the Fourth International Zel'dovich Meeting, Minsk (ICRANET), Belarus (online), 7 September 2020; "Gaussian Process regression of 1D stellar evolution variables", talk at the 15th Würzburg Workshop, Heidelberg, Germany (online), 16 December 2020.

**Isabel Martin and Frauke Gräter:**

Mechanosensing in Focal Adhesions: Pseudokinase Integrin-linked Kinase under Force. Remote BioExcel Summer School on Biomolecular Simulations. Germany, 22-26 June 2020 (poster).

**Wolfgang Müller and Martin Golebiewski:**

"LiSyM SEEK: A FAIR data & models hub and catalogue for LiSyM-Cancer", LiSyM-Cancer Partnering Event (online), 23 April 2020.

**Abraham Muniz-Chicharro, Rebecca C. Wade:**

"Prediction of drug-protein binding kinetics", 12th Annual Colloquium of the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences, Germany (online), 1 December 2020 (poster).

**Ariane Nunes-Alves:**

"Comprehensive characterization of ligand unbinding mechanisms and kinetics for T4 lysozyme mutants using tRAMD", 16th German Conference on Cheminformatics (online), 2-4 November 2020.

**Anna Piras:**

"The Law Of Attraction: Computational Insights into the Role Of Non-Covalent Interactions in Graphene-Based Sensing", Graphene2020 (online), 19-23 October 2020 (poster).  
"Computational Insights into the Role of Non-Covalent Interactions in Graphene-Based Sensing", 1st joint 4EU+/HGS MathComp Annual Colloquium (online), 1-2 December 2020.

**Maria Beatrice Pozzetti:**

"Surface subgroups of semisimple Lie groups", Young Geometric Group Theory IX, Rennes, France, 28 February 2020; "The real spectrum compactification of maximal character varieties", workshop on Simplicial volumes and bounded cohomology, Regensburg, Germany, 22 September 2020.

**Anja Randecker:**

"Coarse geometry of the saddle connection complex" Lost in Translation Surfaces, University of Luxembourg, Luxembourg, 30 January 2020; "What does your average translation surface look like?" Nearly Carbon Neutral Geometric Topology Conference (online), 1 July 2020.

**Benedikt Rennekamp:**

"What's after the break-up? Hybrid simulations of mechanoradicals in collagen", Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany, (virtual seminar), 9 April 2020. "Perspectives on collagen failure", Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany, (virtual seminar), 30 July 2020. "Multi-scale simulations of collagen failure", Matter-to-Life Summer School, online conference, 21-25 September 2020. "Collagen as a buffer of mechanical and chemical stress", Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany, (virtual seminar), 9 Dezember 2020.

**Benedikt Rennekamp and Frauke Gräter:**

"Hybrid Kinetic Monte Carlo / Molecular Dynamics Simulations of bond scissions in proteins", Biophysical Society Meeting, San Diego, USA, 15-19 February 2020 (poster).

**Maja Rey:**

"SABIO-RK Introduction", de.NBI Course "Tools for Systems biology modeling and data exchange: COPASI, CellNetAnalyzer, SABIO-RK, FAIRDOMHub/SEEK" (online), 21 - 23 September 2020.

**Friedrich Röpke:**

"Formation of sdB stars via common envelope ejection by substellar companions", talk at the Annual Meeting of the European Astronomical Society, Leiden, Netherlands (online), 2 July 2020; talk at the Annual Meeting of the German Astronomical Society (online), 25 September 2020.

**Christian Sand:**

"Common-envelope evolution of an AGB star", talk at the Annual Meeting of the European Astronomical Society, Leiden, Netherlands (online), 2 July 2020; talk at the Annual Meeting of the German Astronomical Society (online), 25 September 2020.

**Fabian Schneider:**

"Stellar mergers as the origin of magnetic massive stars", talk at the Annual Meeting of the European Astronomical Society, Leiden, Netherlands (online), 2 July 2020; talk at the Annual Meeting of the German Astronomical Society (online),

25 September 2020. "Supernovae from stripped binary stars", Research Seminar at Monash University, Melbourne, Australia (online), 27 October 2020. "Core Collapse Supernovae from Stripped Stars", 15th Würzburg Workshop, Heidelberg, Germany (online), 14 December 2020.

**Theodoros Soulhanis:**

"Neutron star mergers", talk at the IMPRS retreat, Heidelberg, Germany (online), 25 November 2020; "Neutron star oscillations", talk at the 15th Würzburg Workshop 2020, Heidelberg, Germany (online), 18 December 2020.

**Gabriele Viaggi:**

"Uniform models for random 3-manifolds", Nearly Carbon Neutral Geometric Topology Conference (online), June 2020; "Volumes and random walks on mapping class groups", International young seminar on bounded cohomology and simplicial volume (online), 22 June 2020.

**Eva-Maria Walz:**

"Receiver Operating Characteristic (ROC) Movies", Luminy Conference Mathematical Methods of Modern Statistics 2, Luminy, France (online), 17 June 2020.

**Andreas Weidemann, Ulrike Wittig, Maja Rey, Dorotea Dudas, Wolfgang Müller:**

"SABIO-RK: database for biochemical reaction kinetics", ELIXIR All Hands Meeting 2020 (online), 8-10 June 2020 (poster).

**Ulrike Wittig:**

"Research data management by SEEK/FAIRDOMHub", MIX-UP Kick off Meeting, Brussels, Belgium, 7 February 2020. "Research data management by SEEK/FAIRDOMHub", de.NBI Course "Tools for Systems biology modeling and data exchange: COPASI, CellNetAnalyzer, SABIO-RK, FAIRDOMHub/SEEK" (online), 21-23 September 2020. "SABIO-RK: Curation and Visualization of Reaction Kinetics Data", COMBINE 2020 (Computational Modeling in Biology Network Meeting) (online), 5-9 October 2020. "Research data management by SEEK/FAIRDOMHub", General Meeting of COST Action CA18133 "European Research Network on Signal Transduction" (online), 12-14 October 2020.

**Ulrike Wittig, Stuart Owen, Olga Krebs, Martin Golebiewski, Maja Rey, Alan Williams, Finn Bacall, Xiaoming Hu, Ina Pöhner, Munazah Andrabi, Jacky Snoep, Fateh-meh Zamanzad Ghavidel, Rune Kleppe, Kjell Petersen, Kidane Tekle, Jon Olav Vik, Inge Jonassen, Andrew Millar, Flora D'Anna, Vahid Kiani, Anders Goksøyr, Katy Wolstencroft, Frederik Coppens, Carole Goble, and Wolfgang Müller:**

"FAIRDOM: supporting FAIR data and model management", ELIXIR All Hands Meeting 2020 (online), 8-10 June 2020 (poster).

**Christopher Zapp and Frauke Gräter:**

Mechanoradicals in tensed tendon collagen as new source of oxidative stress, DPG Spring Meeting, Dresden, Germany, 17 March 2020 (poster).

### 9.3 Memberships

**Camilo Aponte-Santamaría:**

Referee for Plos Computational Biology and Nature Communications

**Frauke Gräter:**

Max Planck Fellow of the Max Planck School Matter to Life (since 2019). Fellow of the Marsilius Kolleg, Heidelberg University (2019-2020). Member of the Editorial Board, Biophysical Journal. Member of the Board of Directors, Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University. Associated faculty member, HGS MathComp Graduate School, University of Heidelberg. Faculty member, Hartmut Hoffmann-Berling International Graduate School of Molecular and Cellular Biology (HBIGS), Heidelberg University. Member of the coordinating committee of the excellence cluster "3D Matter Made to Order" (KIT and Heidelberg University). Co-speaker of Biophysics section within the Condensed Matter section of the German Physical Society. Board Member of the Flagship Initiative "Engineering Molecular Systems", Heidelberg University. Referee for Biophysical Journal, Journal of the American Chemical Society, PLoS Journals, Nature Journals, Proceedings of the National Academy of Sciences and German Research Society (DFG).

**Tilmann Gneiting:**

Fellow, European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, United Kingdom. Affiliate Professor, Department of Statistics, University of Washington, Seattle, Washington, United States. Member, Steering Committee, Karlsruhe Institute of Technology Center MathSEE: Mathematics in Sciences, Economics and Engineering. Member, Ensemble Advisory Board, United States COVID-19 Forecast-Hub. Member, Committee on Publications, Institute for Mathematical Statistics.

**Martin Golebiewski:**

Convenor (chair) of the ISO/TC 276 Biotechnology working group 5 "Data Processing and Integration", International Organization for Standardization (ISO). Chair of the project group "FAIR Data Infrastructures for Biomedical Informatics"

of the German Association for Medical Informatics, Biometry and Epidemiology (GMDS). Member of the board of coordinators of COMBINE (Computational Modeling in Biology network). Member of the steering committee of the German National Research Data Infrastructure for Personal Health Data (NFDI4Health). German delegate at the ISO technical committee 276 Biotechnology (ISO/TC 276), International Organization for Standardization (ISO). Member of the national German standardization committee ("Nationaler Arbeitsausschuss") NA 057–06–02 AA Biotechnology, German Institute for Standardization (DIN). Co-chair of the ISO/IEC Joint Ad-hoc Group on Standardization of Genomic Information Compression and Storage (MPEG-G). Member of the steering committee of the AIMe registry for artificial intelligence in biomedical research.

#### Ganna Gryn'ova:

Affiliated junior research group leader: Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University. Member: working group on advancing women's careers in mathematics and computer science, Heidelberg Laureate Forum Foundation (HLFF).

#### Saskia Hekker:

Member of Brite Executive Science Team (BEST); head of an international node of the Stellar Astrophysics Centre (SAC); member of TESS Asteroseismic Science Consortium (TASC) Steering Committee.

#### Sebastian Lerch:

Associate Editor, Monthly Weather Review. Associate Editor, Meteorological Applications. Guest Editor, Nonlinear Processes in Geophysics.

#### Wolfgang Müller:

Member of the Scientific Advisory Board of the BioModels Database. Deputy Chairman of SIG 4 (Infrastructure & data management), German Network for Bioinformatics Infrastructure (de.NBI). Board Member and Treasurer of FAIRDOM e.V.

#### Ariane Nunes-Alves:

Member of the Early Career Board of the Journal of Chemical Information and Modeling.

#### Friedrich Röpke:

Advisory Board: Sterne und Weltraum.

#### 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. Member of scientific committee of the SMPGD (Statistical Methods for Post Genomic Data analysis) workshop series.

#### Michael Strube:

Research Training Group 1994, Adaptive Preparation of Information from Heterogeneous Sources (AIPHES), TU Darmstadt/Heidelberg University/HITS; Associate Editor: Journal of Artificial Intelligence Research.

#### Rebecca C. Wade:

Associate Editor: Journal of Molecular Recognition, PLOS Computational Biology. Section Editor: Molecular Informatics, International Journal of Molecular Sciences. Editorial Board: Advances and Applications in Bioinformatics and Chemistry; BBA General Subjects; Journal of Chemical Information and Modeling; Journal of Computer-aided Molecular Design; Biopolymers; Protein Engineering, Design and Selection. Member of Scientific Advisory Council of the Leibniz-Institut für Molekulare Pharmakologie (FMP), Berlin-Buch. Member of Scientific Advisory Council of the Computational Biology Unit (CBU), University of Bergen, Norway. Member of Scientific Advisory Board of the Max Planck Institute of Biophysics, Frankfurt. Member at Heidelberg University of: HBIGS (Heidelberg Biosciences International Graduate School) 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. Member of Managing Board of Directors, Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University.

#### Anna Wienhard:

Fellow of the American Mathematical Society. Advisory Board, Springer Lecture Notes in Mathematics. Advisory Board, Mathematisches Forschungsinstitut Oberwolfach. Scientific Committee Wissenschaftskommunikation.de. Member Heidelberg Academy of Sciences. Member Berlin-Brandenburg Academy of Sciences and Humanities. Selection Committee, Heinz Maier Leibnitz Preis der Deutschen Forschungsgemeinschaft. Editor Annales Henri Lebesgue. Editor Annales scientifiques de l'Ecole Normale Supérieure. Editor Geometric and Functional Analysis. Editor Geometry & Topology. Editor Proceedings of the London Mathematical Society. Editor Geometriae Dedicata (2011–2020). Editor Forum Mathematicum (2016–2020). Co-Spokesperson: DFG Excellence Cluster 2181 "STRUCTURES: A unifying approach to emergent phenomena in the physical world, mathematics, and complex data", Co-Spokesperson DFG Graduierten-

kolleg 2229 "Asymptotic Invariants and Limits of Groups and Spaces", extended 2020–2024. PI in DFG SFB/TRR 191 "Symplectic Structures in Geometry, Algebra and Dynamics", 2020–2024. PI and Member of Program Committee, DFG Schwerpunktprogramm "Geometry at Infinity" (SPP 2026) 2020 – 2023.

#### Ulrike Wittig:

Member of the STRENDA Commission (Standards for Reporting Enzymology Data).

## 9.4 Contributions to the Scientific Community

### Program Committee Memberships

#### Sucheta Ghosh:

The 34th AAAI Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020. (Program Committee). The 58th Annual Meeting of the Association for Computational Linguistics (ACL), Seattle, USA, 5–10 July 2020. The 21st Annual Conference of the International Speech Communication Association- Interspeech, Shanghai, China, 14–18 September 2020 (Technical Program Committee). The 17th International Conference Applied Computing 2020, Lisbon, Portugal, 18–20 November 2020 (Program Committee).

#### Michael Strube:

Senior Area Chair at the 58th Annual Meeting of the Association for Computational Linguistics, Online, The World, 5–10 July 2020. Area Chair at "The 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics" and the "10th International Joint Conference on Natural Language Processing", Online, The World, December 4–7, 2020. Area Chair at the "28th International Conference on Computational Linguistics", Online, The World, December 8–13, 2020.

### Workshop and Conference Organization

#### Robert Andrassy and Friedrich Röpke:

Members of the Scientific and Local Organizing Committees, 15th Würzburg Workshop, HITS, Heidelberg, Germany (online), 14–18 December 2020.

#### Tilmann Gneiting:

Member, Organizing Committee, MathSEE Symposium 2020: Mathematics in Sciences, Engineering and Economics (online), Karlsruhe Institute of Technology, Germany, 7–9 October 2020.

#### Martin Golebiewski:

Host and chair of the EU-STANDS4PM Workshop "Using patient derived data for *in silico* modelling in personalized medicine", HITS, Heidelberg (Germany), 4 February 2020. Co-host and session chair of HARMONY 2020, EMBL-EBI, Hinxton, Cambridgeshire, UK, 9–13 March 2020. Host and chair of Committee Meetings of ISO/TC 276 Biotechnology working group WG5 "Data Processing and Integration", online, 8 June – 3 July 2020. Co-host and session chair of the COMBINE 2020 Online Forum: 11th Computational Modeling in Biology Network Meeting, online, 5–9 October 2020. Host and chair of the Workshop "FAIR data infrastructures for biomedical communities", 65th Annual Conference of the German Association for Medical Informatics, Biometry and Epidemiology (GMDS), online, 15 October 2020. Task Area Chair "Standards for FAIR Data (TA2)" at the NFDI-4Health Kick-off-Meeting, online, 4–6 November 2020.

#### Olga Krebs:

Best practices in research data management and stewardship. Online course, 19 May 2020 – 3 June 2020. Best practices in research data management and stewardship, Online course organised by ELIXIR Luxembourg and ELIXIR France, 5 – 8 October 2020. Modelling COVID-19 epidemics, online course, co-organized with ELIXIR, ISBE and EOSC, 30 November – 9 December 2020.

#### Ariane Nunes-Alves:

Co-organizer: "LatinXChem" chemistry twitter poster conference, <https://www.latinxchem.org/> online, 7 September 2020. Co-organizer: "The impact of the COVID-19 crisis on women in science: challenges and solutions", EMBL WIS conference, held online <https://www.embl.org/events/covid19-wis/>, 9 September 2020.

#### Friedrich Röpke:

Senatsberichterstatter, Berufungsverfahren Professur Zahnerhaltungskunde, Heidelberg University, Germany, 2020. Member of the Scientific Organizing Committee, session SS5 "New insights of angular momentum transport in stellar interiors", Annual Meeting of the European Astronomical Society, Leiden, Netherlands (online), 1 July 2020.

#### Alexandros Stamatakis:

Organizer of 2020 Computational Molecular Evolution Summer School, Heraklion, Crete, Greece. (The summer school was postponed to 2021).

#### Michael Strube:

Program Co-Chair of CODI 2020, The First Workshop on Computational Approaches to Discourse at EMNLP 2020, Online, The World, 19 November 2020.

**Anna Wienhard:**

Co-organizer of the conference “Teichmüller Theory: Classical, Higher, Super and Quantum”, 5-9 October 2020, CIRM (Centre International de Rencontres Mathématiques), Marseille, France.

**Other contributions****Antonio D’Isanto:**

“A trip across the Galaxy: the search for exoplanets”, Invited lecture for the eight-grade classes of the Milton Hershey School, Pennsylvania, USA, 30 Oct 2020. “B like beyond the human limit or astronomical challenges in the era of AI”. Blog post in the HITS Blog “Via Data” on SciLogs.

**Valentina Disarlo:**

Project Leader in the DFG SPP 2026 Geometry at Infinity for Project 44 “Actions of mapping class group and its subgroups”.

**Isabel Martin and Frauke Gräter:**

“#HITS – Wie Daten Wissen schaffen. Proteine im Crash-test: Von der Achillessehne bis zum Blut.” MAINS Heidelberg, Germany, 13 February 2020.

**Nicholas Michalarakis:**

“A Day in the Life of a Computational Biochemist”; talk at the Summer School REMOTE 2020, “Jugend Präsentiert”, Heidelberg, Germany, (online), 3 July 2020.

**Maria Beatrice Pozzetti:**

Principal Investigator in the RTG 2229: Asymptotic Invariants and Limits of Groups and Spaces. Project Leader in the DFG SPP 2026 Geometry at Infinity for Project 28: Rigidity, deformations and limits of maximal representations and Project 71: Rigidity, deformations and limits of maximal representations II. Leader of a DFG Emmy Noether independent junior research group (Project number: 427903332).

**Anja Randecker:**

Principal Investigator in the RTG 2229: Asymptotic Invariants and Limits of Groups and Spaces. Project Leader in the DFG SPP 2026 Geometry at Infinity for Project 72: Limits of invariants of translation surfaces.

**Anna Wienhard:**

Selected as Henriette-Herz Scout, Alexander von Humboldt Foundation.

**9.5 Awards****Johannes Bracher:**

“ProFID: Probabilistic Real-Time Forecasting of Infectious Diseases”, Young Investigator Group Preparation Program (YIG Prep Pro) award, Karlsruhe Institute of Technology, 2020.

**Frauke Gräter:**

ERC Consolidator Grant, European Research Council, 2020.

**Ganna Gryn’ova:**

Principal Investigator: SFB1249 “N-Heteropolycycles as Functional Materials”.

**Saskia Hekker:**

ERC Consolidator Grant, European Research Council, 2020.

**Sebastian Lerch:**

“Artificial Intelligence for Probabilistic Weather Forecasts”, MINT for the Environment Early Career Research Group award, Vector Foundation, 2020.

**Eugen Rogozinnikov:**

Heidelberger Dissertationspreis Mathematik 2020.

**Carmen Rovi:**

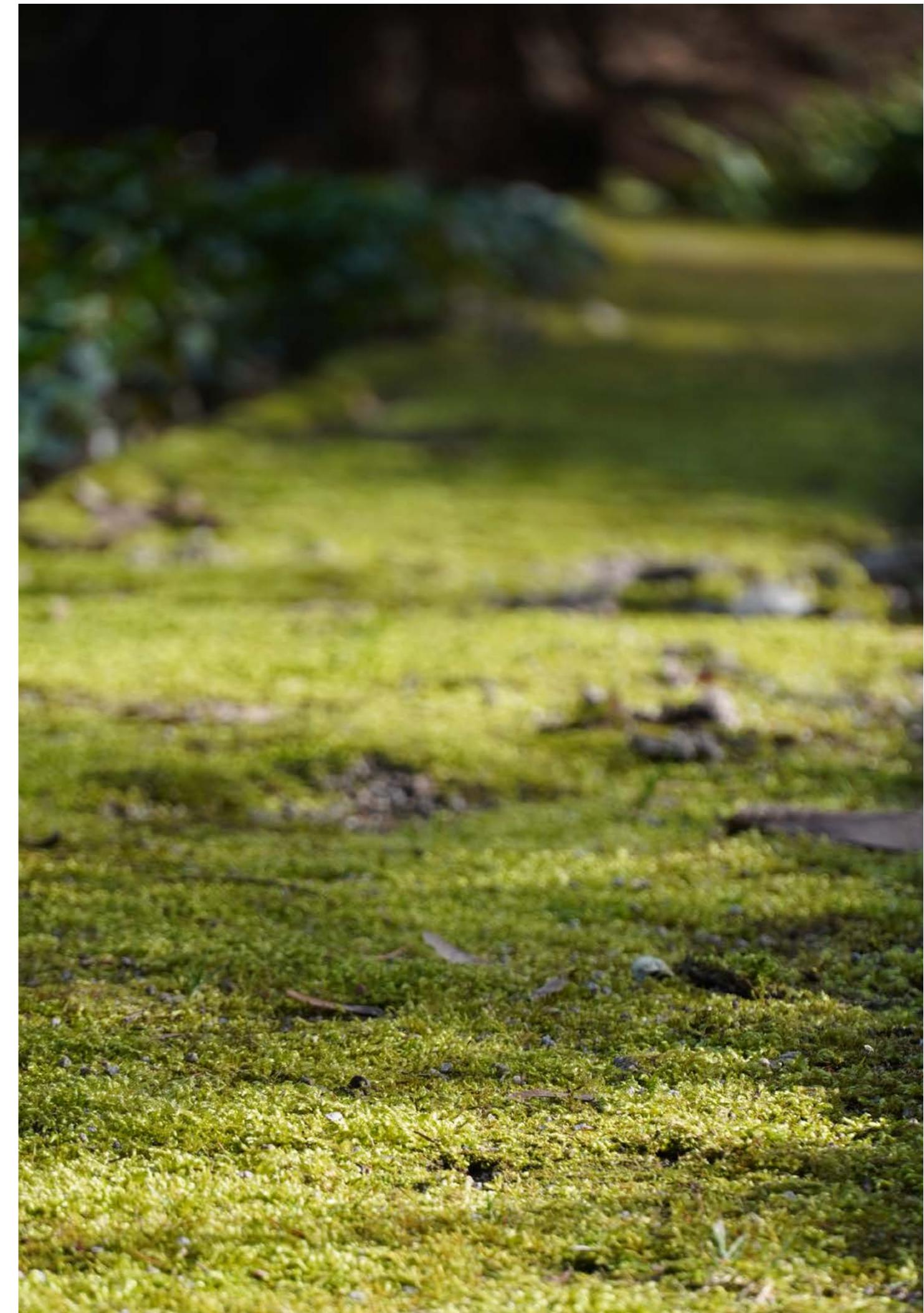
Summer Research in Mathematics Program, MSRI, Berkeley, USA (\$5000 grant and paid accommodation and travel). Invitation from Max Planck Institute for Mathematics, January-June 2020 (comes with significant stipend).

**Fabian Schneider:**

Emmy Noether Fellowship, German Research Foundation, 2020 (declined). ERC Starting Grant, European Research Council, 2020.

**Alexandros Stamatakis:**

Highly Cited Researcher in Cross-Field Research, Clarivate Analytics, 2020.



# 10 Boards and Management



The HITS Scientific Advisory Board. From left to right: Wolfgang Müller (HITS Scientific Director), Tony Hey, Frauke Gräter (HITS Deputy Scientific Director), Alex Szalay, Victoria Stodden, Thomas Lengauer, Adele Goldberg, Barbara Wohlmuth, Dieter Kranzlmüller, Gert-Martin Greuel, Gesa Schönberger (HITS Managing Director), Jeffrey Brock.

## Scientific Advisory Board

The HITS Scientific Advisory Board (SAB) is a group of internationally renowned scientists that supports the management of HITS in various aspects of running, planning, and directing the institute. The SAB is responsible for orchestrating the periodic evaluation of all the research groups of HITS. It presents the results to the HITS management and makes recommendations regarding how to further improve the institute's research performance. In 2020, the board consisted of the following members:

- **Prof. Dr. Jeffrey Brock**, Zhao and Ji Professor of Mathematics at Yale University, USA
- **Prof. Dr. Tony Hey**, Chief Data Scientist, Science and Technology Facilities Council, UK
- **Prof. Dr. Alex Szalay**, Johns Hopkins University, USA
- **Prof. Dr. Victoria Stodden**, School of Information Sciences, University of Illinois at Urbana-Champaign, USA
- **Prof. Dr. Dieter Kranzlmüller**, Ludwig Maximilians University, Munich, Director of the Leibniz Super Computing Center, Germany (Chair, SAB)
- **Prof. Dr. Barbara Wohlmuth**, Chair of Numerical Mathematics at the Technical University of Munich (TUM), Germany.
- **Prof. Dr. Thomas Lengauer**, Max-Planck-Institute for Computer Science, Saarbrücken, Germany
- **Prof. Dr. Gert-Martin Greuel**, University of Kaiserslautern, former Director of the "Mathematisches Forschungszentrum Oberwolfach", Germany

## Shareholders' Board



**Prof. Dr. Wilfried Juling**  
Member of the Board of Directors



**HITS-Stiftung**  
**Prof. Dr.-Ing. Dr. h.c. Andreas Reuter**  
Member of the Board of Directors  
(until May 2020)



**Prof. Dr. Carsten Könneker**  
Member of the Board of Directors  
(since June 2020)  
© Mück/Klaus Tschira Stiftung



**Heidelberg University**  
**Prof. Dr. Jörg Pross**  
Vice-President of Research and Structure  
© Philip Benjamin



**Karlsruhe Institute of Technology (KIT)**  
**Dr. Hanns-Günther Mayer**  
Director of Shareholdings ("Leitung Beteiligungen")

## HITS Management

The HITS Management consists of the Managing Director and the Scientific Director ("Institutssprecher"). The latter is one of the group leaders appointed by the HITS shareholders for a period of two years. The scientific director represents the institute in all scientific matters vis-à-vis cooperation partners and the public.



**Managing Director:**  
**Dr. Gesa Schönberger**



**Scientific Director:**  
**PD Dr. Wolfgang Müller**  
(2019 – 2020)



**Deputy Scientific Director:**  
**Prof. Dr. Frauke Gräter**  
(2019 – 2020)



## HITS

The Heidelberg Institute for Theoretical Studies (HITS) was established in 2010 by the physicist and SAP co-founder Klaus Tschira (1940 – 2015) and the Klaus Tschira Foundation as a private, non-profit research institute. HITS conducts basic research in the natural sciences, mathematics, and computer science, with a focus on the processing, structuring, and analyzing large amounts of complex data and the development of computational methods and software. The research fields range from molecular biology to astrophysics. The shareholders of HITS are the HITS-Stiftung, Heidelberg University, and the Karlsruhe Institute of Technology (KIT). HITS also cooperates with other universities and research institutes and with industrial partners. The base funding of HITS is provided by the HITS Stiftung with funds received from the Klaus Tschira Foundation. The primary external funding agencies are the Federal Ministry of Education and Research (BMBF), the German Research Foundation (DFG), and the European Union.

HITS gGmbH  
Schloss-Wolfsbrunnenweg 35  
D-69118 Heidelberg

**Editor**  
Dr. Peter Saueressig  
Head of Communications

**Contact**  
[info@h-its.org](mailto:info@h-its.org)  
Phone: +49 6221-533 533  
Fax: +49 6221-533 298  
[www.h-its.org](http://www.h-its.org)

Our e-mail addresses have the following structure:  
[Firstname.lastname@h-its.org](mailto:Firstname.lastname@h-its.org)

**Pictures**  
HITS gGmbH (unless otherwise indicated)

All rights reserved. All brand names and product names mentioned in this document are trade names, service marks, trademarks, or registered trademarks of their respective owners. All images are protected by copyright. Although not all are specifically indicated as such, appropriate protective regulations are valid.

**Layout and Design**  
FEUERWASSER | grafik . web . design  
[www.feuerwasser.de](http://www.feuerwasser.de)

ISSN 1438-4159 | © 2021 HITS gGmbH

Twitter: @HITStudies  
Facebook: /HITStudies  
Youtube: /TheHITSters  
LinkedIn: company/the-heidelberg-institute-for-theoretical-studies  
Instagram: the\_hitsters