NanoSpace

NanoSpace DSG Hackathon

Meeting Report

This publication is based upon work from COST Action CA21126 – Carbon molecular nanostructures in space (NanoSpace), supported by COST (European Cooperation in Science and Technology).

COST (European Cooperation in Science and Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation.

www.cost.eu

Funded by the European Union

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

BAU, Istanbul, Türkiye

July 23 — 25, 2025

BAU Bahçeşehir University

DOĞUŞ UNIVERSITY ISTANBUL·1997

COST Action CA21126

Carbon molecular nanostructures in space

(NanoSpace)

NanoSpace DSG Hackathon
July 23 — 25, 2025
BAU, Istanbul, Türkiye

NanoSpace DSG Hackathon

Meeting Report

Published by

Erbahar Research Lab

Organised by

Erbahar Research Lab
Doğuş University
Bahçeşehir University
COST action: CA21126 — NanoSpace

Edited by

Eftal Gezer
Dogan Erbahar

Foreword

One of the main outputs of the COST action NanoSpace (CA21126) is to create a cosmic nanocarbon inventory database to bridging the gap between interdisciplinary scientists. The database efforts are directly subject to WG1, and Database Strategy Group, dedicated especially for the database efforts, carries out this task.

In order to accelerate the database efforts, NanoSpace DSG Hackathon have taken place in July 23 — 25, 2025 at Bahcesehir University, Istanbul, Türkiye with five invited lectures from four invitees, one “Ask an expert” session with six panellists, dedicated working times in a breakout session format, and seventeen participants in total.

With the great synergy refreshed in this event, attendees have expressed their commitment and willingness to continuing the database efforts within NanoSpace.

Prof Dogan Erbahar, WG1 leader
Eftal Gezer, MSc, head of DSG

Organising Committee:
Prof Dogan Erbahar
Prof Lutfi Arda
Eftal Gezer, MSc

Organising Committee

Prof. Dogan Erbahar
Dogus University, Türkiye

Prof. Lutfi Arda
Bahcesehir University, Türkiye

Eftal Gezer, MSc
Gebze Technical University, Türkiye
Dogus University, Türkiye


NanoSpace

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Funded by
the European Union

BAU
Bahçeşehir University

DOĞUŞ ÜNİVERSİTESİ
İSTANBUL·1997

Programme

23 July 2025

Time	Title/Event	Presenter
8:00-9:00	Registration	
9:00-9:40	The ExoMol project: spectroscopic data for hot molecules in exoplanets and elsewhere	Jonathan Tennyson
9:40-10:20	Databases of PAH spectral properties to model astronomical observations: the LAIBrary project	Louan de Bentzmann
10:20-10:35	*Coffee break*	
10:35-11:15	DataBase Systems for Large and Big Data Files for Academic & Scientific Projects	Mustafa Yilmaz
11:15-12:00	Preparation of the day	*Moderators:* Dogan Erbahar & Eftal Gezer
12:00-13:00	*Lunch*	
13:00-14:30	Working time *with flexible breaks and coffee*	
14:30-15:00	Q&A and coordination session	Collaborative
15:00-18:30	Working time *with flexible breaks and coffee*	
18:30-19:00	Progress report session	Participants
19:00-22:00 *(flexible)*	Working time *with flexible breaks and coffee*	

24 July 2025

Time	Title/Event	Presenter
8:00-9:00	Registration	
9:00-9:40	Thoughts on constructing and maintaining databases: molecular data for astrophysics and plasma physics	Jonathan Tennyson
9:40-10:20	Data arrangement and collection for the quantum-chemistry community	Ransel Barzaga Guzmán
10:20-10:35	*Coffee break*	

Time	Title/Event	Presenter
10:35-11:00	Preparation of the day	*Moderators:* Dogan Erbahar & Eftal Gezer
11:00-12:00	Working time	
12:00-13:00	*Lunch*	
13:00-13:30	Q&A and coordination session	Collaborative
13:30-14:30	Ask an expert (hybrid format via Zoom)	*Moderators:* Dogan Erbahar & Eftal Gezer *Panellists:* Jonathan Tennyson, Mustafa Yilmaz, Aníbal García-Hernández, Ransel Barzaga Guzmán, Louan de Bentzmann
14:30-18:30	Working time *with flexible breaks and coffee*	
18:30-19:00	Progress report session	Participants
19:00-22:00 *(flexible)*	Working time *with flexible breaks and coffee*	

25 July 2025

Time	Title/Event	Presenter
8:00-9:00	Registration	
9:00-9:25	Preparation of the day	*Moderators:* Dogan Erbahar & Eftal Gezer
9:25-12:00	Working time *with flexible breaks and coffee*	
12:00-13:00	*Lunch*	
13:00-13:30	Q&A and coordination session	Collaborative
13:30-18:30	Working time *with flexible breaks and coffee*	
18:30-19:00	Progress report session	Participants
19:00-22:00 *(flexible)*	Working time *with flexible breaks and coffee*	

Participants

Invited lecturers

Jonathan Tennyson

Mustafa Yilmaz

Ransel Barzaga Guzmán *(online)*

Louan de Bentzmann

Panellists for “Ask an expert” session

Jonathan Tennyson

Mustafa Yilmaz

Aníbal García-Hernández *(online)*

Ransel Barzaga Guzmán *(online)*

Louan de Bentzmann

Attendees

Lutfi Arda

Muratcan Ates

Sena Celik

Dogan Erbahar

Eftal Gezer

Mete Oztay

Dilara Ickecan

Yusuf Emre Kilic

Eren Umut Ozdemir

Yunus Emre Ozkaya

Mustafa Nuri Öztürk

Hawaiz Mehmood

Contents

Day 1 **1**

Remarks from Jonathan Tennyson's lecture "The ExoMol project: spectroscopic data for hot molecules in exoplanets and elsewhere"..... 1
Remarks from Louan de Bentzmann's lecture "Databases of PAH spectral properties to model astronomical observations: the LAIBrary project"..... 3
Remarks from Mustafa Yilmaz's lecture "DataBase Systems for Large and Big Data Files for Academic & Scientific Projects"..... 4
Summary of the day..... 5

Day 2 **7**

Remarks from Jonathan Tennyson's lecture "Thoughts on constructing and maintaining databases: molecular data for astrophysics and plasma physics"..... 7
Remarks from Ransel Barzaga Guzmán's lecture "Data arrangement and collection for the quantum-chemistry community"..... 7
Remarks from "Ask an expert" session..... 9
Summary of the day..... 9

Day 3 **10**

Summary of the day..... 10

Perspectives **11**

Day 1

Remarks from Jonathan Tennyson's lecture "The ExoMol project: spectroscopic data for hot molecules in exoplanets and elsewhere"

Presented by Jonathan Tennyson, Physics and Astronomy, University College London, United Kingdom



The ExoMol database project begun with a recollection that at its fundamental level Tennyson and coworkers' work that is attempting to solve the Schrödinger equation to calculate the spectroscopic fingerprints needed for the interpretation of hot environments, from exoplanet atmospheres to fusion plasmas. Because 51 Peg b was found in 1995 and known planets became exponentially more numerous (over 6 000 by mid-2025), astronomers have resorted to transit, cross-correlation, and direct-imaging techniques—each with its own merits and issues—to tease out atmospheric signals. JWST infrared, R~3000 transit spectra, imminent ARIEL and Twinkle missions, and Doppler surveys at high resolution from the ground together form a complementary observational arsenal, but all depend on precise molecular data to translate forlorn signals into confident detection and abundance determinations.



To meet this need, ExoMolHD employs first-principies quantum methods—variational nuclear motion codes like Duo and TROVE, ab initio potential-energy and dipole surfaces, and the MARVEL empirical energy-level model—to generate gigantic line lists (up to tens of billions of transitions) with spectroscopic accuracy. These line inventories have already spurred modelling of methane-rich "T4.5" brown dwarfs and water-rich hot Jupiters, well beyond empirical benchmarks such as HITRAN by orders of magnitude in terms of completeness and accuracy.

Beyond line positions and intensities, the group also treats photodissociation and pressure broadening—processes of critical relevance in UV-rich environments and high-pressure planetary atmospheres. Through explicit semi-classical computations, machine-learning programs, and some data from HITRAN and Leiden databases, ExoMol produces temperature- and pressure-dependent cross sections, rate coefficients, and Voigt/Voigt-plus profiles for significant broadeners H_2, He, CO_2, etc. This work extends the database's range of applicability from cold M-dwarf systems to runaway B-star environments and provides retrieval frameworks as well as atmospheric models with information.

All these molecular data products—cross sections, line lists, partition functions, broadening parameters, k-tables, lifetimes, cooling functions, and so on—are provided through an API-based ExoMolHD platform. Optimized for simple integration into retrieval codes like TauREx 3 and for access by laboratory and field scientists without, this living database is updated continuously with new line lists (e.g., PH_3, PN, SO_2) and functionality to support next-generation missions and applications in astrochemistry, remote sensing, combustion, and climate science.

Behind it all is an engaged international community—from the early work by Yurchenko,

Tennyson, and co-workers in the development of variational spectroscopy to recent progress in photodissociation (Pezzella), predissociation (Mitev), and pressure broadening (Guest, Buldyreva, Gamache). Their collaborative effort puts ExoMolHD at the very leading edge of molecular data science, paving the way for discoveries in astronomy, atmospheric physics, and beyond.

Remarks from Louan de Bentzmann's lecture "Databases of PAH spectral properties to model astronomical observations: the LAIBrary project"

Presented by Louan de Bentzmann, IRAP, France
Co-contributed by Christine Joblin, Giacomo Mulas, Karine Demyk, Jean-Michel Glorian, Axel Pérignon, Vincent Marty



The explanation for the astronomical Aromatic Infrared Bands (AIBs) in terms of polycyclic

aromatic hydrocarbons (PAHs) is a challenging task. de Bentzmann et. al highlights how excitation conditions are varied—especially UV radiation fields—and PAH charge state, chemical variability, and local density produce complex emission spectra that cannot be explained by simple harmonic models. Laboratory experiments on isolated PAHs reveal that, upon the absorption of a UV-photon, these molecules heat up to temperatures of up to 2000 K and then cool down in sequences of IR emissions; such "hot PAH" photophysics yields broadened, shifted bands whose anharmonic nature must be treated to match observations.

To fill in these modelling gaps, the team reviews existing analysis tools—PAHfit for the extraction of dust features and pyPAHdb for observation fits to harmonic spectra—before introducing the LAIBrary project's Monte Carlo-based approach. By combining quantum-chemistry databases of vibrational frequencies and intensities with stochastic calculations of dissociation rates, UV absorption, IR emission, and fluorescence, they generate synthetic IR spectra across temperature ranges, inserting empirical anharmonicity functions to reconcile band broadenings and shifts found in real data. The CosmicPAHfit toolchain then deconstructs observed or calculated spectra into constituent bands, extrapolates temperature-dependent anharmonic parameters via polynomial fitting, and outputs simply interpretable files—enabling reproducible figure generation and speedy sharing of spectral information.

In the team's approach, big spectral files reside on the outside with SQL table asynchronous references to raw spectra, processed band profiles, species metadata, instrument data, and experimental contexts (phase, sample conditions, institution). The CosmicPAH IRDB is hosted in a Flask backend with Jinja2 templates, with a JavaScript frontend offering interactive access; containerization offers portability and ease of deployment. In summary, the authors note that by bringing together experimental measurements and theoretical simulations under a community-based repository, the LAIBrary project will enable astronomers and chemists to have the anharmonic PAH spectra required to unravel the infrared signatures of the universe.

Remarks from Mustafa Yilmaz's lecture "DataBase Systems for Large and Big Data Files for Academic & Scientific Projects"

Presented by Mustafa Yilmaz, Department of Engineering Physics, Gaziantep University, Turkey



Mustafa Yilmaz outlines the necessities to develop an open-access, multi-disciplinary

nanocarbon database for both current researchers and future missions. It begins by stating the need for a centralised database where laboratory measurement data, theoretical simulation data, and observational astronomy data converge on a searchable, extensible platform for efficient screening, modelling, and mining of nano-carbon species.

Under the hood, the infrastructure needs enterprise-grade Linux servers—Oracle Linux, RHEL, ALMA, Rocky, Ubuntu—protected by firewalls and built out into application, database, and storage layers with backup and load-balancing systems to offer reliability and security. Such hardware platform and OS supports both MySQL and PostgreSQL, the former appreciated for its simplicity, fast performance, and pluggable storage engines, and the latter for its conformant SQL behaviour, extensibility (e.g. materialised views, triggers, richer data types), and robust transactional guarantees. Yilmaz outlines the choice of PostgreSQL is clear and incisive.

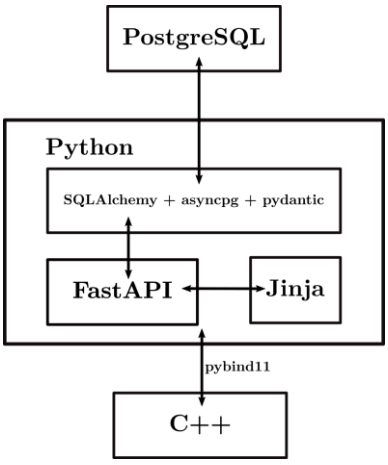
In order to make the system remain administratively friendly, a collection of graphical utilities for PostgreSQL—pgAdmin, DBeaver, DataGrip, Navicat, OmniDB—provides graphical interfaces for all from basic schema browsing to advanced debugging, data visualisation, and cloud integration, dramatically reducing the learning curve and making administration easy. Both administrators and end users appreciate built-in dashboards, keyboard shortcuts, and visual editors that speed up frequent tasks and minimise mistakes.

Finally, based on the key gaps identified by the NanoSpace members previously—missing metadata, patchy data coverage, unreliable APIs, and poor UI on some legacy databases—Yilmaz outlines setting a direction for community-funded improvements. Priorities should include enriching spectral and material constants for many nanocarbon types, normalising formats for easy ingest, and establishing a "meta-database" layer that consolidates atomised resources into an integrated, versioned, citable platform.

Summary of the day

The programme has begun with an introductory presentation and acknowledgement on NanoSpace, by WG1 leader Dogan Erbahar.

After the first day's talks, the technical backbone, database structure, so-far progress on back-end, and the hackathon strategy are explained to the attendees by DSG head Eftal Gezer.



The attendees are divided into four breakout groups with a collaborative discussion as explained in the following table, and have added to the hackathon team on GitHub.

Breakout Group 1 (BG1)	**Breakout Group 2 (BG2)**	**Breakout Group 3 (BG3)**	**Breakout Group 4 (BG4)**
Front-end design and implementation	**Processing of FITS files spectrum and spectral lines DB pipeline**	**Implementation of JS9 web tool for FITS files**	**Development and implementation of data entry and administration dashboard**
Yusuf Emre Kilic	Sena Celik	Yunus Emre Ozkaya	Eren Umut Ozdemir
Muratcan Ates	Dilara Ickecan		Mustafa Nuri Ozturk

BG1 determined a strategy to design the front-end with `tailwindcss` package. They have drafted a design for the homepage, and the molecule record page.

BG2 determined a strategy to process the FITS files and pipelining the spectral lines from relevant DBs by `astropy`, `astroquery`, `specutils` Python modules. They have begun by creating a pipeline for Splatalogue.

BG3 began to examine the JS9 technical structure.

BG4 determined a strategy to design the front-end and back-end backbone of the panel. The dashboard will consist of AdminLTE dashboard template, React TypeScript library. Form validations will be implemented by Formik for React library. API requests will be handled by Axios. The structure will be modular to achieve a clear and readable code. Security of the system will be handled by AES256 encryption and user tokens will be stored in `LocalStorage`.

Day 2

Remarks from Jonathan Tennyson's lecture "Thoughts on constructing and maintaining databases: molecular data for astrophysics and plasma physics"

Presented by Jonathan Tennyson, Physics and Astronomy, University College London, United Kingdom

Jonathan Tennyson defines his expertise on building and evolving a family of molecular databases—HITRAN, GEISA, ExoMol and its offspring (ExoMolOP, ExoMolHR, ExoPhoto, ExoAtom, LiDB), and broader portals like QBD, LXCat, KIDA, BASECOL, Phys4Entry and VAMDC. Tennyson points out that a successful database is contingent on good data quality, ready access, stable metadata, robust user interaction, flexible formats, prompt support, and ongoing maintenance. These principles have shaped each project's design from early HITRAN releases decades ago to the current ExoMol community.

Founded in 2011 by Tennyson and Yurchenko, ExoMol was created with the goal of providing detailed, high-temperature molecular line lists for the simulation of hot atmospheres and exoplanets. By solving the molecular Schrödinger equation and gathering billions of transitions of spectroscopic quality, ExoMol broadens the scope and precision of legacy compilations. Its compact "ExoMol format", Python-oriented post-processing scripts (ExoCross), and public API facilitate seamless integration into retrieval sets and modelling workflows, with four-yearly database updates and occasional conference releases ensuring continuing relevance.

In addition to line lists, ExoMol provides temperature- and pressure-dependent cross sections, partition functions, pressure-broadening coefficients, opacity k-tables, radiative lifetimes, cooling functions, Landé g-factors and oriented-dipole data. Specialist extensions—ExoMolHR for empirical high-resolution spectra, ExoPhoto for photodissociation cross sections, and ExoAtom for atomic transitions—extend its application to fusion plasmas, laboratory spectroscopy and beyond. Other datasets such as NASA polynomial specific heats and LiDB's vibrational lifetimes add to the resource.

Tennyson also places ExoMol within a larger network of secondary data products—Quantemol's plasma services for trade, KIDA's kinetic rates, BASECOL's collision cross sections, Phys4Entry's entry physics and VAMDC's portal for federated metadata—demonstrating how a variety of user communities can coexist by leveraging common infrastructure. By being led by best practices in data curation, standards and user support, ExoMol and its friends demonstrate how sustainable scientific databases depend as much on technological innovation as on a firm commitment to their user communities.

Remarks from Ransel Barzaga Guzmán's lecture "Data arrangement and collection for the quantum-chemistry community"

Presented by Ransel Barzaga Guzmán, University of La Laguna, Spain

The urgent need for improved systems to arrange, store, and share data within the quantum chemistry research community is emphasized. Quantum chemistry inherently involves highly

complex computational problems, particularly the many-body problem, where computational costs rise steeply with system size. Simulations of molecular, interfacial, and bulk systems can demand hundreds to thousands of CPU hours and generate extremely large files—often in the hundreds of megabytes—containing millions of lines of text. These files are frequently formatted with non-standardized keywords, making automated parsing and processing challenging. The diversity of quantum chemistry codes, each employing its own unique input and output structures, exacerbates these issues, creating significant obstacles for reproducible research, code interoperability, and the adoption of standardized application programming interfaces.

The current landscape features several important initiatives aimed at addressing these challenges. Established repositories such as the Protein Data Bank, the Computational Chemistry Comparison and Benchmark Database, and the Cambridge Crystallographic Data Centre facilitate reproducibility, yet they do not fully resolve interoperability issues between different computational codes. More specialized projects, such as Quixote, have introduced structured formats like Chemistry Markup Language and RDF metadata, while the MolSSI QCArchive project offers a suite of interoperable, Python-based tools—including QCElemental, QCEngine, QCFractal, QCPortal, and QCSchema—that employ HDF5 formats to standardize the storage and retrieval of quantum chemistry data. TREXIO provides a robust, open-access, HDF5-based format tailored specifically to quantum chemical properties, complete with API support and an emphasis on long-term data transferability. In the educational domain, the eChem project uses Jupyter notebooks as an interactive teaching tool for quantum chemistry and computational methods. Additional enabling technologies, such as the Atomic Simulation Environment, facilitate handling and visualization of multiple file formats, while HDF5's hierarchical structure allows for efficient management of large, heterogeneous datasets with associated metadata.

Despite these advances, key challenges remain unresolved. There is an increasing demand for the integration of machine learning and deep learning models to produce chemically robust and adaptable AI systems. However, the creation of high-quality training datasets, especially for systems with intricate electronic structures, is resource intensive. Another persistent difficulty lies in selecting optimal molecular descriptors to enhance model performance and generalizability. Addressing these issues will require not only algorithmic innovation but also rigorous data standardization.

Within this context, the DSG initiative aims to develop a realistic database prototype capable of consolidating all information produced by the COST Action Nanospace, with full accessibility for project members by December 2026. This effort involves navigating API licensing constraints in collaboration with Spanish institutions, managing the complexity of SQL-based queries to ensure database performance, and determining whether a Python-only architecture can effectively serve the diverse needs of experimental, observational, and theoretical research pipelines. Sustainable database maintenance is also identified as a critical factor for long-term success.

In conclusion, while Jupyter notebooks offer a flexible and widely accessible environment for the presentation and analysis of quantum chemistry data, their reliance on evolving software dependencies introduces potential stability risks. The combination of Python-based tools with robust data formats such as HDF5 offers a promising path forward for efficient data handling,

conversion, and interoperability. Success in this endeavour will depend on careful definition of database variables, adherence to standardized formats, and a long-term vision for accessibility and maintainability across the quantum chemistry community.

Remarks from the "Ask an expert" session

Panel discussion with Jonathan Tennyson, Mustafa Yilmaz, Aníbal García-Hernández, Ransel Barzaga Guzmán, Louan de Bentzmann

During the "Ask an expert" session, it is discussed that amongst the spectral line databases CDMS, Kurtz, NIST Atomic Lines are pioneer. For the scientific part, additional data scope for DIB spectra, experimental lab spectra are discussed. In the technical part, experimental lab spectra are stored in JDX file format.

The necessity for the scientific starting point must be theoretical and observational part by considering the shareholders' scientific expertise.

Summary of the day

All working groups worked in a naturally-developed full coordination. All breakout groups uttered their willingness to participate the database efforts after the hackathon. All breakout groups have continued coding. BG3 have developed the entry point for JS9 package with minor deficiencies.

On the basis of "Ask an Expert" session, the scientific priority is determined as observation spectra, reaction kinetics, and DFT & MD. For the spectral line pipeline, it is determined that the priority should be CDMS, JPL, NIST Atomic Lines database.

Since the high majority of the attendees are affiliated with the Dogus University, it is decided to continue the work at Erbahar Research Lab at Dogus University's Dudullu Campus on the Day 3 by the request from the attendees.

Day 3

Summary of the day

The day have begun with the Mustafa Yilmaz's visit to Erbahar Research Lab at Dogus University's Dudullu Campus.

Meanwhile, breakout groups have continued coding in coordination to correct the deficiencies. High levels of enthusiasm, eagerness, dedication, and passion are observed on the breakout groups. It is decided to do weekly online meetings on every Thursday evening after the hackathon.



Perspectives

By wrapping out the invited lectures, “Ask an expert” session, and the inter-breakout group discussions several unequivocal patterns have been settled in order to guide future evolution of the NanoSpace database project. To begin with, the hackathon brought to the forefront that robust, on-premises server backbones—adorned with dedicated system administrators, enterprise-class hardware, and secure firewalls—are still essential. Without being given a rock-solid, secure footing, even the most stunning software designs will collapse.

No less important is constraining the scientific scope. While NanoSpace's frontiers extend from theoreticians, experimentalists, and observers, it is exactly that scope that risks weakening influence so far. In other words, the Action’s broad scientific scope and the expertise of the database efforts’ shareholders are double-edged sword. In the beginning, the primary effort must be centred on those areas where our volunteers are most experienced: observational astronomy and computational physics; specifically, observational spectra, reaction kinetics, and DFT & MD. By keeping early development centred around these two pillars—building validated, high-accuracy molecular spectra ab initio, and seamlessly integrating them with telescope data pipelines—we can rapidly create value and gain traction.

The hackathon left no doubt that metadata and interoperability are the building blocks of any viable platform. Beyond raw data files, our database must support provenance, and instrument context; it must have a clear database interface that federates existing resources into an integrated, citable system. Building metadata standards and light-weight APIs up front will unlock downstream benefits—simple data ingestion, reproducible analysis, and enhanced cross-project collaboration. In these three domains—infrastructure, scientific focus, and metadata governance—NanoSpace's long-term success is laid.

After those starting points, the scope should expand into data mining and machine-learning tools, but those need more experts.