How to Better Leverage Data to Unblock Pharma R&D

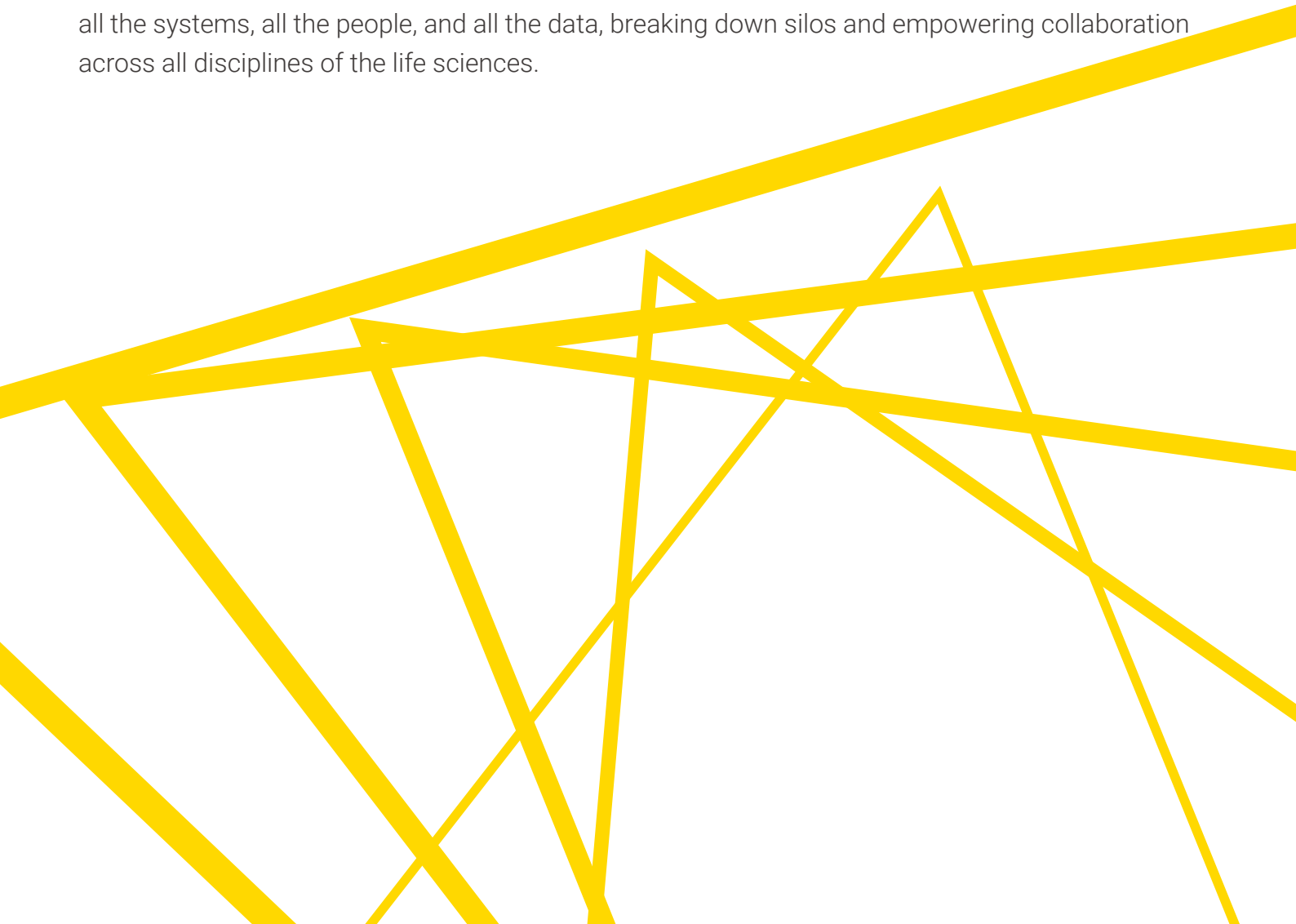
Accelerating Drug Discovery with Collaborative Data Analysis

To date, 62% of the global population has been vaccinated against COVID-19, with vaccines that were only made possible through unprecedented international collaboration. Successful drug discovery relies on the collaborative efforts of multidisciplinary teams sharing research data not just across departments, but across countries, specializations, industries, and (importantly) with external organizations and regulators. Many pharma companies have established consortiums and programs with academia to facilitate such collaboration. Unfortunately, work in siloed tools is a barrier to successful collaboration.

Cheminformaticians work in Python and R, which can only be understood by other scripters. Lab managers work with lab information management systems, which are not always understood by technicians' electronic lab notebooks. Computational chemists develop models on compound activity, which are black boxes to synthetic chemists.

These barriers prevent teams from collaborating efficiently, achieving their true potential, and scaling their efforts. Siloed data, tools, and teams turn data collection, curation, and analysis into time-consuming, inefficient processes. Siloing of tools in particular has led to Powerpoint and Excel becoming essential tools for scientific communication, which is a perverse state of affairs.

One way to lower these collaboration barriers is to make sure everyone can access and analyze the data and focus their effort where they can provide the most value. Low-code/no-code tools connect all the systems, all the people, and all the data, breaking down silos and empowering collaboration across all disciplines of the life sciences.



Integrate all relevant data from any source

Data integration is a dense process that requires strong domain knowledge and proficiency in data wrangling. Lab managers, biologists, pharmacologists, and synthetic chemists spend up to 80% of their time wrestling with data. From drug target data to omics[1] data, from SMILES[2] to molecular property descriptors, from in vitro assays to ADMET[3] data, or treatment-response and trial-design data. Typically collected from biological, pharmacological, and biomedical datasets, this data is large in size and complex, consisting of multiple data types.

What's needed is a single consistent environment to access, transform, and connect all required data.

Wave Life Sciences, an interdisciplinary biotechnology company based in the US, faced this issue when they had to integrate multiple data types and sources, languages, and APIs for bio/cheminformatic services.

---

1. The term "omics" refers to a number of areas of study in biology, all of which end in the suffix -omics: Genomic, transcriptomic, proteomic, metabolomic.

2. Simplified Molecular-Input Line-Entry System is a specification in the form of a line notation to describe the structure of chemical species using short ASCII strings.

3. Chemical absorption, distribution, metabolism, excretion, and toxicity data. This data is frequently used to determine whether a compound is suitable to proceed to the clinical stage.

Experiment data submission reduced from weeks to minutes at Wave Life Sciences

The company used KNIME's open low-code/no-code analytics platform to access multiple data sources and types to provide curated data used in experiment data submission. By doing so, they reduced the time teams needed to submit experiment data from weeks to minutes.

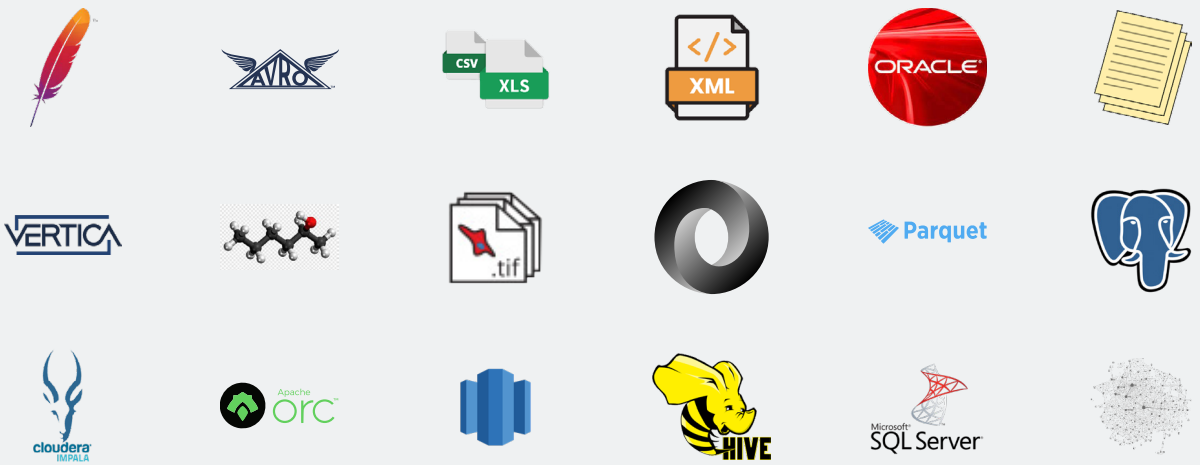
> "Scalable, repeatable, and reusable ETL processes are saving data scientists an enormous number of steps. What took an entire week, now – with curated data – takes only 5 minutes," Kenneth Longo, Head of Discovery Data Science, at Wave Life Sciences.

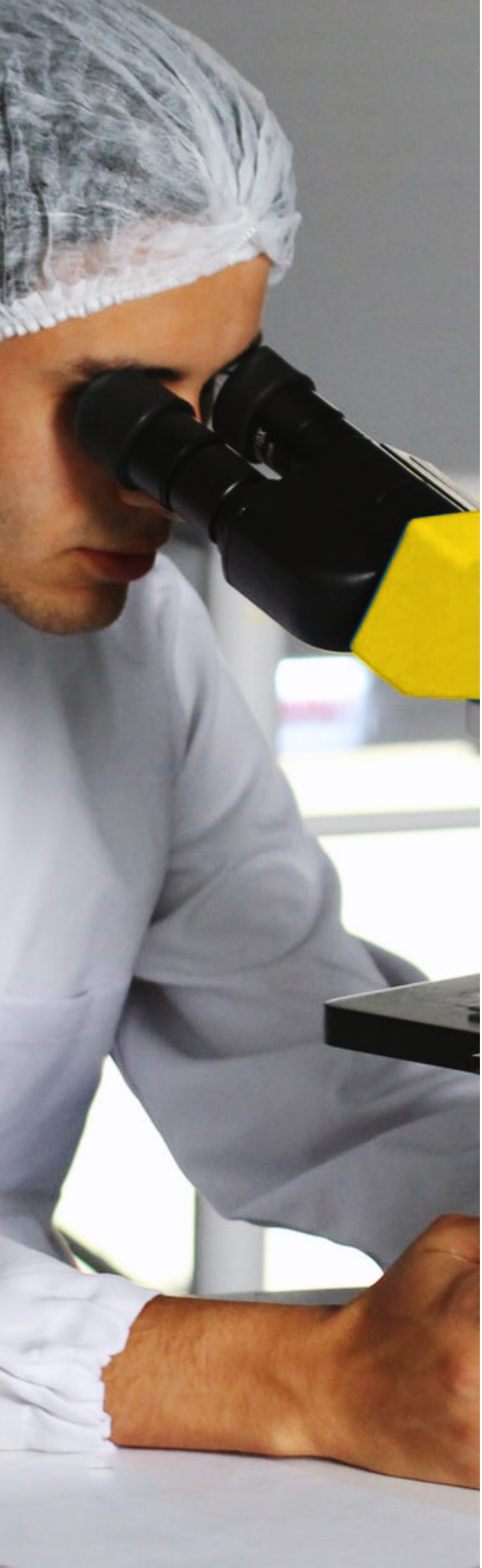
The entire team of 20 multi-disciplinary professionals, including 8 high-level domain experts can collaborate using the same solution.

The team integrates R, Python, Perl, and command line tools. They access multiple APIs and merge different data types and locations — databases, data lakes, on-prem, and in the cloud.

> "And we have no more discussions about how we're challenged with ETL or transition issues – because it's just smooth," said Longo.

Some of the data types and sources that need to be accessed and integrated in drug discovery





Expand analysis and include non-AI experts

In drug discovery, advanced machine learning (ML) and artificial intelligence (AI) techniques can be used effectively for everything from reading and analyzing literature to predicting how a drug will eventually perform during clinical phases. The advantage is the ability to quickly carry out extremely sophisticated pattern recognition on huge datasets. AI/ML techniques have the potential to improve drug success rates and lower costs by as much as 70%, as reported by **Insider Intelligence**.

The crux is that multidisciplinary teams are not always able to use these novel, highly advanced methods without becoming AI experts first. Instead ML engineers and specialized data scientists are hired to develop AI-based algorithms. While they support the team, they lack the relevant domain expertise.

In a challenge to predict drug development outcomes, **Novartis scientists collaborated with MIT data scientists**. They used state-of-the-art machine learning algorithms and leveraged the valuable expertise in drug discovery of Novartis scientists, statisticians, portfolio managers, and researchers. The winning teams developed predictive models that significantly outperformed existing models. The challenge clearly showed the additional value of enabling lab scientists and data scientists to work together.

CENTOGENE accelerates early drug discovery with faster biomarker investigation

To really improve drug success rates and achieve lower cost goals, teams need to draw on the expertise of a broader pool of participants. All the stakeholders – the biologists, pharmacologists, chemists, and data scientists – need access to advanced analytics.

As a leading generator of rare disease insights, CENTOGENE wanted to enable their biomarker department to use highly advanced algorithms for biomarker investigation. Widely used at every stage of drug discovery and development, biomarker identification has the potential to optimize the drug discovery, development, and approval processes.

Using low-code/no-code tools like KNIME, data scientists can deploy browser-based data apps without needing to bring in IT or acquire knowledge of front-end programming languages. With the right level of guidance and interaction, the lab scientists can then leverage advanced analytics techniques through the data app and inject their own domain expertise where necessary.

"Guided analytics" is an approach that enables teams to combine their expertise and build applications that give each subject matter expert – the experimental chemist or the synthetic biologist for example – access to just the right amount of guidance and interaction during the analysis.

CENTOGENE wanted a new solution that would give them the flexibility to accommodate the frictionless integration of new or improved algorithms. Their new solution implements guided analytics. It gives multidisciplinary teams access to advanced algorithms and enables non-AI experts to explore data independently.

With guided analytics, computational models can be built and enriched with input from the domain experts — for example, to narrow down the field of new drug candidates, or at CENTOGENE, new biomarkers.

KNIME has enabled CENTOGENE to collaborate easily and accelerate the identification of biomarkers.

> "There is less manual work for the AI experts, and the non-AI experts can interact with intuitive visualizations to get even more out of the data than was previously possible."
>
> Anne Schwenk, IT Project Manager Artificial Intelligence, CENTOGENE

Collaborate with coders and non-coders

In drug discovery, the work is often carried out through collaboration with multiple research groups around the world. In a typical scenario, experimental chemists are sending lists of compounds to computational chemists, who run predictions and send back the results. Every time the experimental chemist questions these results, the computational chemist has to take time out of their day to provide explanations, which slows down collaboration.

Computational chemists are frequently using programming languages, such as Python, to write advanced scripts that control data flow, transformation, and analysis and set up computational prediction models. These scripts tend to grow as the experiment progresses, and are seldom well-documented. The steps are not inherently clear, which makes them difficult to understand, discuss, and reuse.

To improve the work between experimental chemists and computational chemists, for example, an environment is needed that supports mutual understanding.

How Nuvisan improved sharing and reusing Python functionality in predictive modeling

One advantage of the no-code, low-code interface is that it provides a common ground for collaboration. As the computational chemist builds a predictive model, the visual programming environment provides an intuitive visualization of how the data is flowing. Now, when the experimental chemist questions the prediction model results, the computational chemist can simply share their work. The visual workflow provides immediate clarity, making it easier to understand, discuss, and reuse.

But what if the computational chemist wants to continue using Python? Team productivity is higher if everyone can work with their preferred tools.

Nuvisan is a pharmaceutical company working in drug discovery and development. The scientists wanted to enable teams to share their work across the entire organization. A lot of this work, coded in Python, was only available to other coders.The "Science CRO" Nuvisan Innovation Campus Berlin developed a suite of workflows for processing, analyzing, and predictive modeling of chemical data that could be shared across the organization.

One of these workflows calculates properties such as molecular weight and solubility by using a variety of chemistry integrations and implementing ML models. It includes a molecule "Standardizer" which ensures all the compounds are properly represented in the database.

> Dora Šribar, Molecular Modeler at Nuvisan, developed the Standardizer by integrating Python scripts inside the Nuvisan solution. "For each Python node you can specify different Python environments which gives the flexibility to transition smoothly from one machine to another," she said.

Python users often code their solutions by creating their own custom libraries. These custom libraries transform and analyze data according to very specific requirements. A custom Python library requires a specific configuration in order to run. But this specific configuration and the related dependencies make it difficult to share and reuse. If you don't have the exact same configuration on your machine, the custom library won't run.

By bundling the Python script into the Standardizer with KNIME, Šribarcan can share this functionality with colleagues who don't know Python. She can link coders and non-coders: The Standardizer can be easily copied into other teams' workflows and used on different machines. Propagation support ensures that any package dependencies inside the Python scripts are included in the Standardizer.

Reuse data processing functionality and reduce manual work

The sheer volume of data that comes with the rise in the use of digital technologies and cloud platforms is impossible to manage effectively with conventional tools. High Throughput Screening (HTS) is a case in point.

Typically, analysis of these large sets of raw data is conducted with spreadsheet-based tools, with the work split across multiple isolated files. Working in individual spreadsheets means that labs are constantly defining a process for how to work with the data. After cleaning and filtering, they are left with just a table of results and no historical record of what happened to the data in that table. The process is lost, and the next time it has to be repeated.

The SciLifeLab, a world-leading collaboration of four universities in Sweden and Scandinavia (Karolinska Institutet, KTH Royal Institute of Technology, Stockholm University, and Uppsala University), wanted to find an alternative to assess HTS data more efficiently.

SciLifeLab increases efficiency of raw HTS data assessment with reusable workflows

Using KNIME, SciLifeLab has built reusable workflows for high throughput screening, data analysis, processing, and hit identification. The workflows automatically import multiple spreadsheets and merge the data. Labs are no longer limited to the amount of data they can process. Advanced data science processing techniques deliver screening results much faster. And these workflows can be easily shared across teams.

Each step of the workflow is automatically documented. In addition, explanatory notes can be inserted underneath each step of the process, and longer annotations can be added around groups of functionality. The sequence of operations is easy to follow, and each step can be individually inspected.

> "The motivation for this study was to help laboratories access all kinds of raw data generated from HTS."
>
> Jordi Carreras-Puigvert, Assistant Professor, **Karolinska Institutet**

Share knowledge automatically by integrating deployment

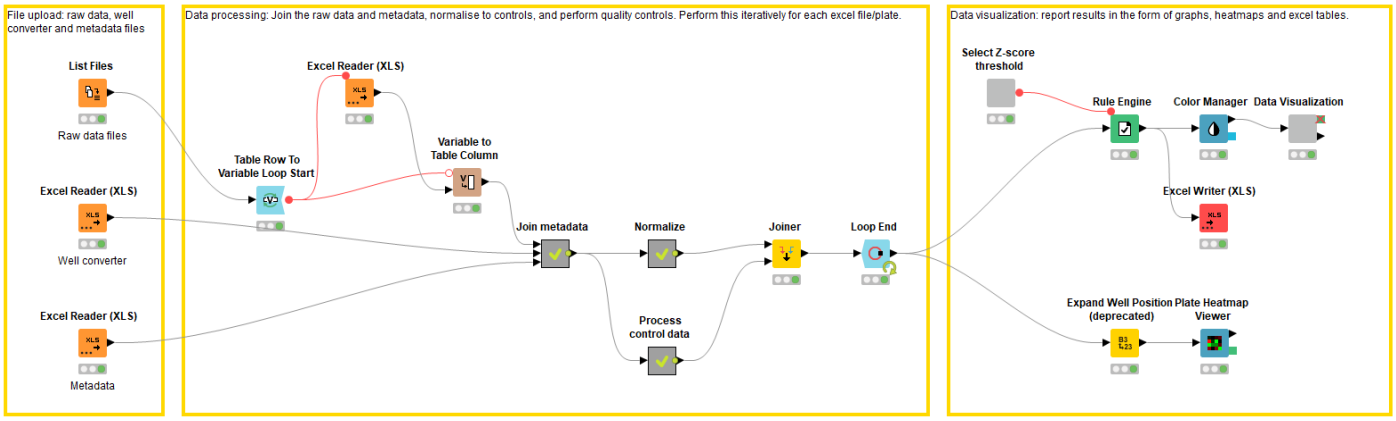
Computational scientists know the drill: Prepare data, train and optimize models with different machine learning methods, evaluate performance, and select the best model. Next, these predictions need to be shared across teams. More often than not, there is no easy way to do this. The trained model has to be saved to a file and all the preprocessing has to be copied to the new environment, which takes time and increases the risk of mistakes.

Using the **integrated deployment** approach, the preprocessing protocol and the best prediction model can be automatically deployed into a production workflow. A simple browser-based application can now call this model prediction workflow to share the work.

This type of approach would, for example, enable a computational scientist to automatically replicate results. In discussions about which compounds to proceed with, the computational scientist can share the app with the synthetic chemist. The synthetic chemist can independently examine model performance, and with this new knowledge review compound selection and prioritization.

Unblock collaboration with no-code/ low-code analytics

With KNIME's visual programming environment, multidisciplinary teams can easily build and share analytics apps with intuitive drag-and-drop, regardless of role or experience with data analytics tools. The visual approach also makes it easier to share analytic concepts.



In comparison to purely commercial no-code/low-code tools, KNIME, as an open-source platform, lowers the barrier to entry. With it, thousands of people in your organization could download it and start building solutions immediately.

Open Means Collaborative

Open platforms fuel cooperation on innovation. They are built by and for collaborative problem solving. KNIME makes it especially easy to share and collectively develop new tools and insights. This particularly helps distributed research teams join forces on complex analytics workflows across departments, regions, and hierarchies, and lets them tap into the cumulative know-how of external expert networks to solve internal data problems.

Drug discovery companies using KNIME

Chiesi
People and ideas for innovation in healthcare

ALEXION

CENTOGENE
THE RARE DISEASE COMPANY

WAVE
LIFE SCIENCES

Vernalis
RESEARCH

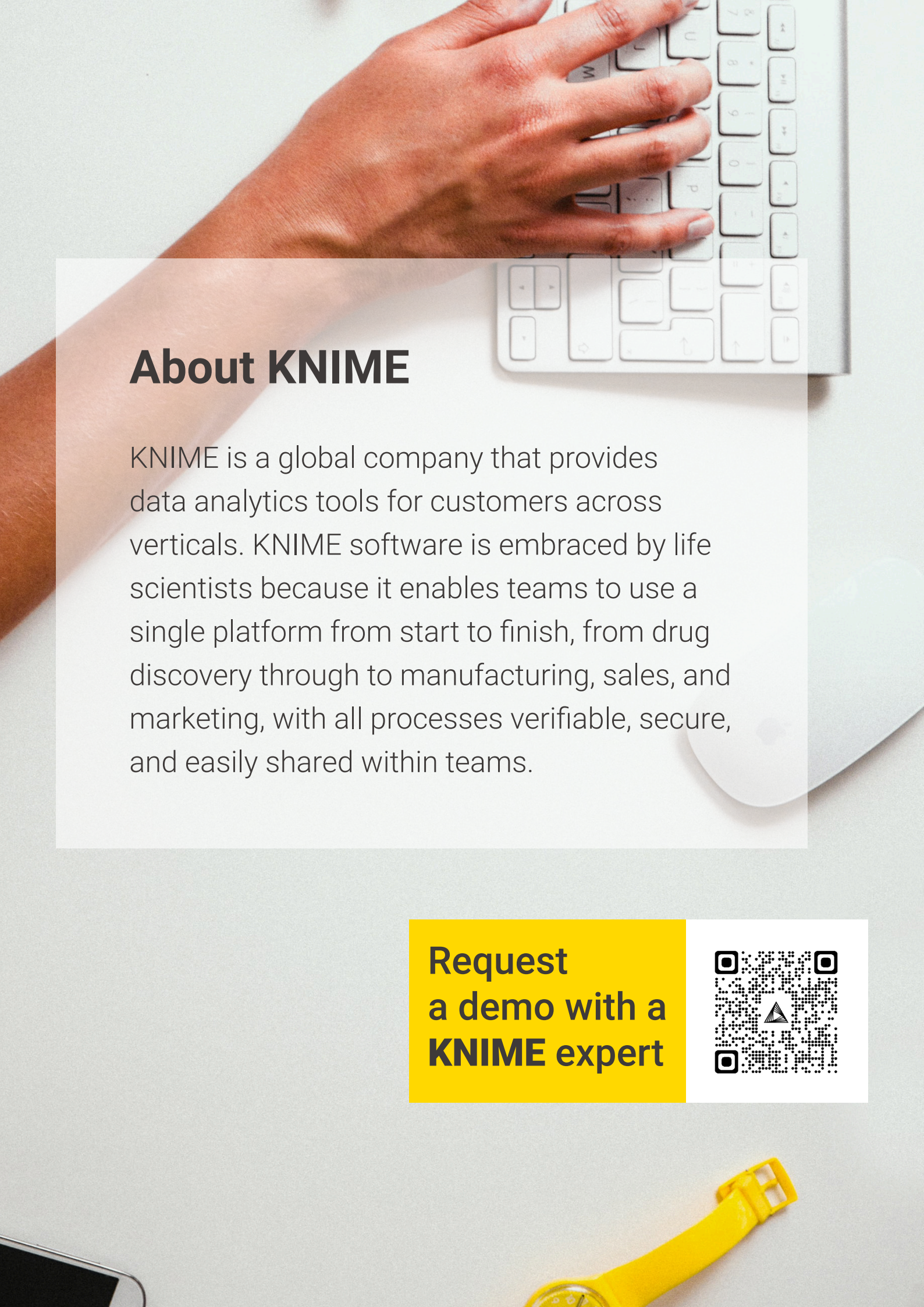
Why KNIME for Collaborative Drug Discovery

1. **Extensible and flexible**: The integrative platform enables both data and domain experts to access any data type from any data source. Universal connectivity is ensured, with access to 300+ data sources and integrations with all the relevant tools and environments.with all the relevant tools and environments.

2. **Visual programming environment**: The analytics process is intuitively visualized, enabling data analysis concepts to be shared easily across multidisciplinary teams.

3. **End-to-end-coverage**: Sharing work is easy through automatic deployment to a report or browser-based data app. With Integrated Deployment, there is no need to piece together different production solutions.

4. **Open source is futureproof**: Community-driven innovation gives teams access to the latest analytic techniques that commercial vendor-driven platforms can't keep pace with.



About KNIME

KNIME is a global company that provides data analytics tools for customers across verticals. KNIME software is embraced by life scientists because it enables teams to use a single platform from start to finish, from drug discovery through to manufacturing, sales, and marketing, with all processes verifiable, secure, and easily shared within teams.

Request a demo with a KNIME expert