[Interview] Expediting Drug Discovery Through Advanced Machine Learning

Sept. 18, 2020 by Andrii Buvailo

The application of next-generation data analytics tools, powered by machine learning and artificial intelligence (AI) components, has become a long-term strategic priority for most companies in the pharmaceutical and biotech industries. However, such systems have to make sure the organisational data is findable, accessible, interoperable, and reusable across different sub-systems, applications, departments, teams, and even companies.

Aigenpulse, a technology company at the forefront of data management and analytics in the Life Science industry, has built a portfolio of tools for working with organisational research data at scale and accelerating the discovery and development of better targets and candidates using advanced machine learning technologies.

I asked several questions to Dr. Satnam Surae, Chief Product Officer at Aigenpulse to learn some specifics about the company and its capabilities, and how it can help Life Science organisations build data-centric research workflows. Satnam has been active in the Life Sciences for more than 10 years. While originally focusing on Biochemistry, he discovered early on his passion for applying information technologies to biological challenges.

Andrii: Satnam, can you briefly outline the company's story, what is Aigenpulse's key goal? Who are the founders and how it all became what it is today?

Satnam: Aigenpulse was founded by Tobias Kloepper, PhD, to help accelerate drug development processes through enabling the use of modern data technologies in Life Science companies. Researchers use the Aigenpulse Platform to structure, share and analyse their research data, and scale repetitive tasks, eliminating the frustrations of dealing with Big Data and accelerating discovery and development of better targets and candidates using advanced machine learning technologies. Our key goal is to streamline drug discovery and help our customers to become the best they can be.

Today's life sciences organisations will generate data from different experiments and may use tens or hundreds of different IT and informatics systems across different laboratories, departments and locations, but it's unlikely that these platforms will all speak the same language, or output data in a common format. Proprietary software will likely result in the creation of silos of data/information that isn't easily evaluated in context with other data. Aigenpulse considers this to be unacceptable in the age of FAIR principles. We are committed to ensuring that data on our Platform can be exploited as the lifeblood of the organisation, and that means providing the ability to interact with and exchange data across systems, sources and software.

Aigenpulse Ltd was founded in May 2016, with an early focus on Proteomics data processing which soon extended to other data types including transcriptomics, assays, cytometry and more. Earlier this year, we launched the latest version of the Aigenpulse Platform for importing, integrating, processing, linking, visualising, exploring and analysing biological data across multiple sources and types into a common frame of reference. It enables users to safely leverage all their data assets to create new insights, build predictive foresight and share findings with stakeholders.

Experiment Suites, which solve specific challenges and are focused on one type of scientific data, are mapped to organisation-wide sets of samples, vocabularies and ontologies, enabling a centralised, accessible, auditable repository of data (raw, processed and analysed), analysis, ML models and reports.

We have also recently launched the CytoML Experiment Suite – an automated, end-to-end, machine-learning solution specifically aimed at enabling streamlining and automation of cytometry analysis at scale. With it, users benefit from a single point-of-truth about all cytometry data across an enterprise organisation.

Andrii: Your team develops a data analytics platform for the life sciences - what exactly does it do? Can you outline several typical use cases of how pharma/biotech clients can benefit from using your platform?

Satnam: Built specifically for the R&D enterprise, the Aigenpulse Platform unifies silos promoting data re-use■, provides automated processing, analysis and report templating, offers in-built statistics,

visualisation and machine learning tools■, enabling high-quality outputs to be generated at multiple stages of the R&D life cycle.■

A recent client, a mid-sized, well-funded and fast-growth biotech company, was creating more data than they could handle with their established methodology. The struggle hindered identification of essential research insights. The head of proteomics already had a lot of internal data. Coming in all shapes and sizes, the heterogeneous datasets had to be re-structured. With our help they were able to process existing data in raw and processed forms, and then import this freshly organised information. The client also wanted to analyse existing public data alongside their own. We integrated multiple external data assets into the Platform, including datasets from ProteomeXchange, TCGA, GTEx, TRON and more. This allowed the client to generate data-driven conclusions from internal and external datasets.

After the import, all data was coherently structured and accessible via an easy-to-use web interface and API. The Aigenpulse Platform became the most used research software across the client's organisation. With our tools, their experts now have an easy way to add, track and analyse new datasets.

When another client came to Aigenpulse for help with in-silico target validation, their established analytics routine took 3-4 weeks per target. Their data exploration was hampered by external data and analytics tools, which couldn't be combined efficiently with in-house data assets. Without the ability to rapidly analyse large data assets, the client could not drive important development decisions. Ineffective in-silico target validation kept slowing them down, and every additional data parameter slowed them down even more. Since our client already had a lot of in-house data and relied on external datasets, we imported these assets into the Aigenpulse Platform. To maximise data handling efficiency, analytics were also integrated into the client's package.

Once everything was accessible via the unified interface and API, we started to optimise the analytics workflow. The performance tuning data structures and machine learning improved efficiency and scalability of reliable complex analytics procedures. The fully automated analytics made in-silico target validation as rapid as possible. Now, the Aigenpulse Platform offers real time data analysis and generates reports upon every data import. The Aigenpulse Platform integration cut down the data analysis time frame from 3-4 weeks to real time. The fully automated in-silico validation routine increases data processing efficiency and saves time. In this case, our client needed the Aigenpulse Platform to identify targets for biologics development. Now they have a tool to efficiently choose and progress validated leads, without wasting time on lengthy analytics or losing promising signals across large datasets.

Andrii: What's the core technology behind the analytics platform? What kind of machine learning approach are you using, what data types are compatible with the platform?

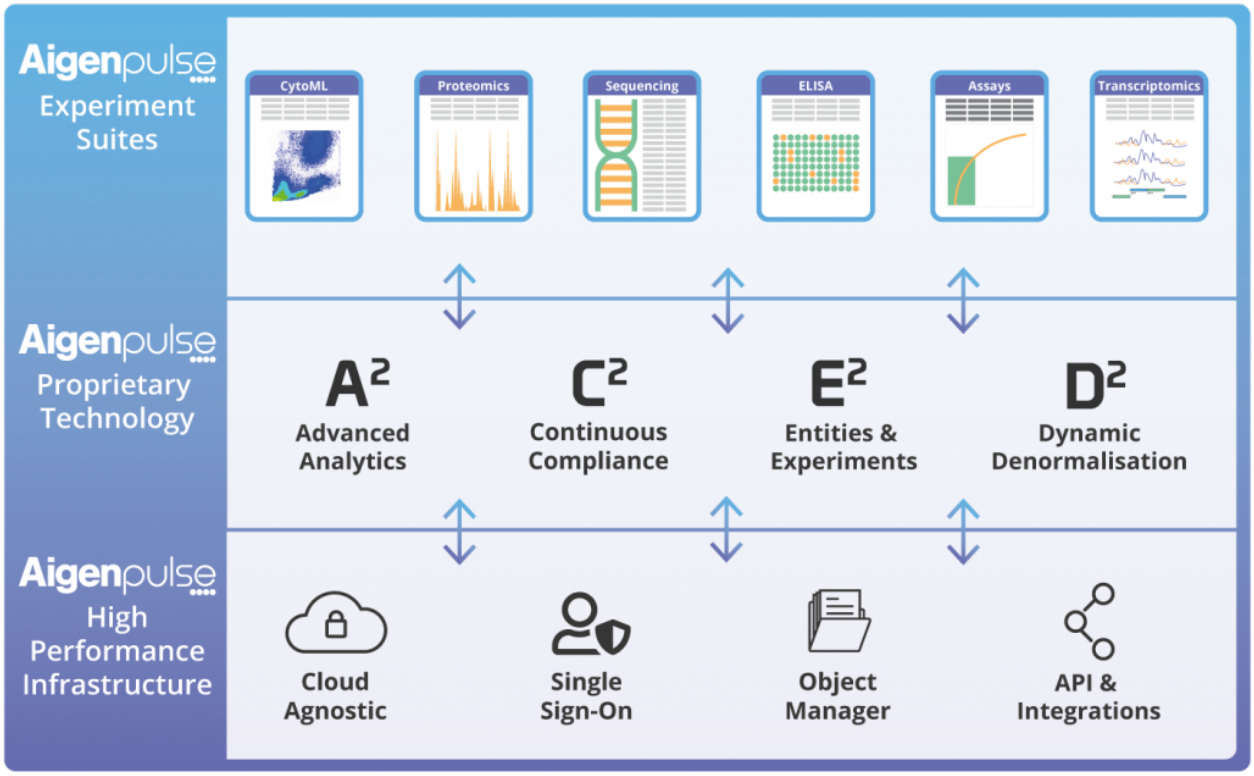
Satnam: We are developing a coherent data analytics backend system to enable the analysis of complex data assets at the click of a button. This enables data to more efficiently guide scientists to develop a more holistic and informed process and ultimately, to develop new drugs more quickly and efficiently.

Aigenpulse A2 (Advanced Analytics) is an extensible framework which encapsulates all the machinery, database models, definitions and forms required to run various analytical tasks using a vast array of data science packages, data processing pipelines and proprietary algorithms. It allows for scalability with large numbers of users and large, complex datasets and the ability for granular process reproducibility (aligning with regulatory compliance and pharma FAIR principles).

Aigenpulse C2 is our proprietary technology which automates and encapsulates all testing and evidence generation processes (including our continuous integration / continuous deployment pipeline). This ensures that during development cycles the Aigenpulse Platform is tested and retested in accordance with GAMP 5 V-model structures.

Aigenpulse E2 is our proprietary technology for building automatically connected data from across different Aigenpulse Experiment Suites. The Aigenpulse E2 technology was built in response to the observation that scientific data is produced in experiments (genomics, proteomics, cytometry, assays, etc) to generate values about specific sets of biological entities, such as genes, proteins, peptides, cells, tissues, biological samples.

Aigenpulse D2 is our technical data modelling technology that provides configurable functionality for the user in how they explore and interrogate their data through the web front-end, e.g. pivoting datasets on particular purification antibody, or drug, or indication and showing persistent statistics. This puts real power in the hands of users.



Andrii: The market of AI-powered data analytics platforms in the Life Sciences is becoming a crowded place, with already more than 230 biotech-focused companies developing or applying some kind of machine learning-based analytics or prediction models in drug discovery and clinical research. What are the key competitive differentiators of your product that make it a unique value proposition?

Satnam: There is no other enterprise data analytics SaaS solution suited to the life sciences quite like the Aigenpulse Platform. Our clearly differentiated approach is focussed around three key customer value drivers: reuse, quality and efficiency.

With the Aigenpulse Platform, scientists can process hundreds of datasets simultaneously and at scale, freeing them up for higher value tasks. The platform can easily integrate with ELNs and LIMSs, in-house data lakes, for a single-point-of-truth for sample/experiment meta-data, and public data sources, such as TRON,TCGA, and GTeX.

Recognising the significant regulatory requirements of life science organisations, we developed an automated system to collect, template and store the evidence required for any configuration of the Platform. It offers automatic QA/QC on datasets, an end-to-end breadcrumb trail to ensure reproducibility of analytics and GxP alignment for use in regulated development, clinic and manufacturing.

The Aigenpulse Platform is cloud-agnostic and can be deployed as a single-tenant Platform on AWS, GCP, Azure or Private Cloud. There is an annual subscription for named users and Experiment Suites. Aigenpulse provides an ITIL-compliant support level agreement, complete GXP validation when required, a 24/7 helpdesk and full online documentation.

Andrii: Do you have any projects related to COVID-19 research? If so, can you explain how your product is contributing to a global fight against coronavirus?

Satnam: In collaboration with Guy's and St Thomas' NHS Foundation Trust and the Francis Crick Institute in London, and the European Bioinformatics Institute (EMBL-EBI) in Cambridge, UK, King's College London (KCL) launched the Covid-IP (Covid–ImmunoPhenotype) project in March 2020, to better understand the immunophenotype of patients infected with SARS-CoV-2 (the coronavirus responsible for Covid-19).

The Covid-IP project performed immunophenotyping on blood samples from >120 Covid-19 patients, consisting of eight complementary flow cytometry panels per patient to capture the major populations of immune cells and rare populations, as well as activation markers. This generated thousands of FCS files, requiring significant manpower to organise the analysis, which is manual. Individual scientists had to create each gate and process each file individually. In a study of this scale and urgency, this created a dual challenge of manpower and the possibility of variability in the data due to the manual gating.

As a direct comparison to the manual gating that was carried out on this project, automated pipelines were implemented for flow cytometry analysis using the CytoML Suite in the Aigenpulse Platform. It automated all steps from data import, QC, gating, statistical analysis and visualisation. This enabled researchers to apply guided algorithms to mimic human gating strategies to the entire dataset without manual intervention.

Compared with the KCL manual pipeline, automated processing provided a strong correlation (Pearson's R = 0.93) and reduced variation for each gating step. The fast processing time reduced the full time equivalent (FTE) from >10 over eight weeks using manual gating, to 1.5 over two weeks using automated

gating.

CytoML removes the barrier to data processing, automating the mundane, routine tasks of gating and allowing scientists to quickly process data and focus their time on data exploration and assessments and testing their hypotheses. It allows the sharing of gated cytometry data between researchers working across different platforms, making it an invaluable tool for validating and verifying the reproducibility of analyses.

- Aigenpulse