Programme Thesis

Hypersensory Intelligence: Olfactory Perception

v1.0

Claire Donoghue, Programme Director

CONTEXT

This document presents the core thesis underpinning a programme that is currently in development at ARIA. We share an early formulation and invite you to provide feedback to help us refine our thinking.

This is not a funding opportunity, but in most cases will lead to one – sign up here to learn about any funding opportunities derived or adapted from this programme formulation.

An ARIA programme seeks to unlock a scientific or technical capability that

+ changes the perception of what's possible or valuable
+ has the potential to catalyse massive social and economic returns
+ is unlikely to be achieved without ARIA's intervention.

PROGRAMME THESIS, SIMPLY STATED

An overview of the programme thesis, accessible & simply stated

Artificial intelligence (AI) models can describe a photograph, compose a symphony and predict the structure of proteins but still cannot smell whether the milk in your fridge has gone off. Olfaction is entirely missing from these models, even though volatile chemicals in the air carry information about disease, food safety, contamination, ecology, and more. Biological systems have been decoding these signals through olfaction for a billion years.

Our programme goal is a general-purpose olfactory perception system, matching or exceeding the sensitivity of biological systems on defined challenge tasks. Imagine a phone that smells a disease flare days before symptoms appear in the millions of patients struggling with chronic conditions [1,2], or prevents the $1 trillion in annual food loss caused by spoilage [3]. If the chemical world becomes computationally legible, the impact will be transformative.

The recipe for AI's success has been the same across fields: standardised sensors, large open datasets, and learned representations. Human consumption of photos and music drove standardised sensors that led to unified datasets and applications. We don't have this driving force for olfaction. Powerful sensing options exist, from analytical lab equipment to deployable gas sensor arrays, but the field is fragmented, and without shared infrastructure these efforts produce point solutions rather than general capability [4–7]. We propose a unified approach: build a high-resolution cross-domain dataset, learn the discriminative features of olfactory space from it, and use those features to build a general-purpose sensor. Recent work suggests the discriminative signal can be learnt directly from data [8,9].

This programme will fund both the missing data infrastructure and the teams who will use it to deliver early breakthroughs. We will build a standardised hardware platform and a cross-domain dataset spanning health, food, and the environment. Alongside these efforts, we will fund teams learning the language of scent through new representation models, and teams translating those representations into cutting-edge, deployable sensing capability.

This programme thesis is derived from the ARIA opportunity space: Extending Our Perception

PROGRAMME THESIS, EXPLAINED

A detailed description of the programme thesis, presented for constructive feedback

Volatile chemical signals carry information relevant to some of the most pressing challenges across health, food, and the environment. This programme aims to give computers a sense of smell and advance machine olfactory perception towards the maturity of computer vision, so that olfaction becomes a general-purpose sensing capability that can be integrated into multimodal models. Achieving a general-purpose olfactory perception system requires standardised sensing capability that can be integrated into applications, alongside datasets that enable new algorithmic capabilities for representation learning.

Biological olfaction achieves remarkable performance on a modest sensory bandwidth, dogs outperform humans in tasks such as diagnosing cancer, detecting allergens, finding mould in buildings, identifying explosives, and locating survivors through meters of rubble [10–12] despite receiving far less raw information per sniff than their visual system delivers per glance.

Our **North Star** is for olfactory perception to become a ubiquitous capability that reaches the maturity of computer vision, such that it can be readily deployed in frontier AI models, robotics, and industrial applications.

Imagine a general-purpose sensing system, as ubiquitous as the camera and integrated into our devices. Pressing challenges from pre-symptom disease flare detection to over 1 billion tonnes of yearly food waste [3] demand digital solutions that are able to access the world of volatile chemicals. If we can learn the most significant features of olfactory space this becomes possible.

In support of ubiquitous digital olfaction, the **goal of the programme,** over 3-5 years, will be to build a general-purpose artificial olfactory perception system. The system should be able to address at least 3 applications with a common hardware platform using the learned representations of olfactory space. Candidate applications could include longitudinal health signals, food spoilage, and food safety.

We define an ***olfactory perception system*** as a hardware platform, algorithmic component, and dataset contribution to produce representations of olfactory space from acquired signals.

The goal of the programme will be metricised based on the following sub-goals

Programme Goal:

To build a general-purpose artificial olfactory perception system, with

1. **Cross-domain generality:** One system that can achieve generality across at least 3 distinct Volatile Organic Compounds (VOC) application domains without application specific hardware by leveraging learned representations.
2. **Biological parity:** The system matches or exceeds sensitivity of the current best benchmark. This benchmark could be derived from other systems such as a biological or analytical benchmark (e.g. trained canines or analytical chemistry equipment).
3. **Publicly accessible resources:** The hardware, datasets, and AI models are publicly accessible and demonstrated to be usable by several independent research groups.

Moving towards the North Star

Long term, the capability to **deploy miniaturised solutions outside of the lab** is vital. We believe the current priority is capturing high-resolution data and the solutions should not initially be constrained by miniaturisation or cost-reduction. By achieving the programme sub-goals, market forces will drive lower-cost, miniaturised deployments after programme completion. Nonetheless, we are open to funding projects to explore capabilities that open the pathway towards widespread deployability.

Why it's worth shooting for

General-Purpose Olfactory System

The value of a general-purpose olfactory system spans beyond the specific applications we demonstrate within the programme. A general olfactory 'sense' would enable novel challenges to be addressed much more rapidly and allow integration of olfaction into multimodal systems alongside other modalities.

This programme is designed to demonstrate the feasibility of transferring olfactory representation knowledge between domains by selecting three applications. The three application domains are selected to test whether learned representations transfer between them. If they do, the same sensor platform can address new applications without being redesigned. To demonstrate the underlying value of the technology as a platform, we review the value of some applications.

Human health

Over 4000 Volatile Organic Compounds (VOCs) have been mapped [13] in human breath, sebum, and other bodily fluids. Olfactory perception offers an opportunity for non-invasive longitudinal monitoring of metabolomic signatures. Diagnosis via olfactory sensing has deep roots in clinical medicine. Physicians from Hippocrates onwards used scent as a diagnostic tool [14], and paramedics are still trained to recognise the fruity breath of diabetic ketoacidosis [15]. Despite centuries of clinical observation and decades of research into electronic noses, no clinically validated computational tool for diagnosis via olfactory sensing exists at the point of care [5,16]. However, there is substantial evidence that disease alters the chemical profile of breath, skin, and bodily fluids in ways that are detectable by biological systems [10,12,15]. Trained dogs can identify multiple cancer types with sensitivity exceeding 90% in some controlled studies [10,12], and have demonstrated the ability to detect infectious diseases including COVID-19 with comparable accuracy [11]. The examples are not limited to dogs' perception. Joy Milne, who has hereditary hyperosmia, noticed a change in her husband's scent years before his Parkinson's diagnosis. The finding was validated through mass spectrometry of sebum biomarkers that now underpins an experimental skin swab test with 90% accuracy [17,18].

VOCs produced by changes in cellular metabolism, inflammation, and oxidative stress have been identified as potential biomarkers across a wide range of conditions, including lung, breast, prostate, colorectal, ovarian, and bladder cancers, Parkinson's disease, diabetes, asthma, Chronic Obstructive Pulmonary Disease (COPD), tuberculosis, malaria, *C. difficile* infection, chronic kidney disease, liver cirrhosis, cystic fibrosis, inflammatory bowel disease, and pre-eclampsia [1,2,4,19–21].

We expect the outcomes of this programme to advance the following application areas

- Human health signals: analysing volatile biomarkers over extended periods to forecast future clinical events and track disease progression. Additionally, we will focus on estimating critical blood-based health indicators using minimally invasive methods, reducing the reliance on traditional venous needle draws. We have picked these two applications as we have noted promise in longitudinal studies [1,2,20] where cross-sectional studies have struggled and because blood biomarker prediction holds advantages over disease prediction, which is more subject to diagnostic label noise.
- Indoor environmental monitoring for human health and the exposome. We spend the majority of our time inside and our health is largely affected by the quality of indoor air (mold, pests, endocrine disruptors, carcinogens, and pathogens). Mapping VOCs in indoor environments could unlock significant public health benefits for the prevention of disease.

Beyond health

Beyond health applications, there is an opportunity to collect large datasets in adjacent domains much more rapidly. Many of the signals in these domains are likely to connect to microbiome or microbial VOC analysis, which strengthens the case for cross-domain transfer. VOCs play an important role across the breadth of our food production and provisioning ecosystem. The ability to better monitor agricultural assets from livestock to crops, all of which are living systems with distinct VOC profiles, offers economic and social benefits. Olfactory sensing also offers value for detecting trace allergens, and predicting and preventing food spoilage, and detecting food safety risks such as *Salmonella,* which produces hydrogen sulfide and certain aldehydes.

Food spoilage, caused by moulds, yeast, and bacteria, is an enormous problem, ⅕ of the world's food is wasted and ⅓ of people are lacking food, in addition food waste has an economic impact of $1T a year in waste and contributes to up to 10% of global greenhouse gas emissions [3]. Of the 1.05 billion tonnes of food wasted annually, 60% is discarded by consumers at home. The remaining 40% is lost further up the supply chain, where real-time detection could help supermarkets and distributors prioritise perishable stock before it spoils. Detection at each level could replace blunt processes like expiry dates set months in advance with real-time information, whether that means extending the shelf-life of safe stock or deciding which pallet to distribute first.

The applications of a general purpose olfactory system span beyond what will be tested in the programme. The goal is to test specific high value applications and to demonstrate an ability to learn a representation that spans between those applications. If successful, the programme would catalyse applications that are not in scope initially through the learned representation and general-purpose hardware. Those follow-on applications might include ecological systems for which the importance of olfaction has long been recognised by

natural historians and scientists [22,23]. The potential applications span beyond biological applications. In industrial settings, olfactory perception could detect the off-gassing of a machine as an early warning signal to predict machine failure prior to a fault within a critical manufacturing process or to predict reduction in quality of a manufacturing process without having to send a sample to an analytical chemistry lab, which is time consuming. We have selected biologically adjacent applications where volatile signatures are likely to share common origins, particularly where microbial metabolic processes are a primary source of VOCs [24].

Why it's differentiated

There was a burst of investment and activity in digital olfaction hardware following the creation of the first "e-nose" in the 1980s [25], but the landscape remains deeply fractured and underfunded relative to other sensory modalities.

Major initiatives, such as the DARPA Real Nose programme in the early 2000s [26], drove early innovation but were ultimately hindered by overpromising and underdelivering, particularly about generalisation and in-the-wild usage. As a result, subsequent efforts have focused on point solutions due to a lingering negative view. More recently, targeted pushes from the NSF [27] and NIH [7] have funded specific health and air quality monitoring applications.

The commercial landscape for volatile chemical sensing remains fragmented and has not progressed towards general-purpose olfactory perception. Each sensor technology has different sensitivity, selectivity, and drift characteristics, so every new application requires its own hardware-algorithm pairing. Data collected on one platform cannot be compared with data from another. This incompatibility persists even with analytical lab equipment, due to a lack of calibration [4]. Companies use sensor arrays for food quality and livestock monitoring [6] , or analytical chemistry platforms for breath diagnostics, each optimised for a specific application. The market rewards solving one problem well, with no commercial incentive to build shared infrastructure that would benefit the whole field.

We are not the first to identify the need for data to drive representation learning: the NSF Convergence Accelerator Workshop concluded that the field needs large standardised datasets, open benchmarks, and shared hardware platforms [28]. A recent review of breath VOC biomarker development found no biomarkers had reached clinical validation, primarily due to fragmented protocols and irreproducible data across studies [4]. However, no programme has acted on this diagnosis at scale.

Some existing olfactory data resources do already exist. These datasets will inspire and have utility within the programme, however no large scale open dataset of VOCs paired with

application-relevant labels that cuts across disciplines exists. The existing datasets fall into three categories.

- Molecular databases such as mVOC [29], HMDB [30] and FooDB [31] aggregate compound identities and species associations from published literature, but provide neither standardised raw signals nor application relevant outcome labels.
- In-the-wild datasets such as SMELLNET [32] and New York Smells [33] pair sensor readings with object or scene derived labels from images for specific categories.
- Biomarker studies are typically small and fragmented and difficult to collate [4]. Commercial breath analysis services demonstrate that standardised VOC analysis is technically feasible [34].

A notable representation-learning effort, the DREAM Olfaction Prediction Challenge [35] targeted a different goal, predicting how molecules smell to humans. They proposed the models could be used to create new fragrances and better understand human olfactory percepts. In this programme, we seek to create digital olfaction where the learnt percepts are targeted towards a set of functional applications rather than mimicking human perception. This valuable initiative helps to demonstrate the potential to learn a perceptual map of the underlying data.

We propose a programme that builds the world's first cross-domain VOC dataset using standardised hardware capable of capturing rich signals including a range of sensors and lab grade analytical chemistry equipment, e.g. Gas Chromatography-Mass Spectrometry (GC-MS). The dataset will be used to discover transferable representations of olfactory space, discover reproducible signals from that data and use those representations to specify and demonstrate deployable general-purpose sensing capabilities. The value of running these workstreams in parallel is that the learned representations can inform what data to collect next, while hardware teams can test whether their sensors resolve the chemical dimensions that the representations identify as discriminative. The programme requires each element to co-evolve the system. The target will initially span biologically-aligned applications including health, food, and environmental applications, to test the transferability of learned olfactory representations on adjacent challenges and to inform deployable sensor platform design.

Why now

We believe this thesis is tractable *now* due to a confluence of recent advances:

1. **There is new strong evidence of low-dimensional structure in VOC space**
   In 2020, it was shown that 21 physicochemical features (measurable molecular properties) could be used to learn olfactory metamers (a scent mixture with no shared molecules that smell identical) [8]. In 2023, the Principal Odour Map demonstrated that a GNN embedding of molecular structure generalises across multiple human olfactory perception tasks [36], and the Principal Odour Map was

extended in 2025 to include mixtures [37]. While the precise dimensionality remains debated [38], these advances reinforce the biological existence proof of low-dimensional sensory space: even a fruit fly with just ~50 classes of receptors can achieve robust, complex odour discrimination [39].

2. **Cross-domain transfer has a mechanistic basis.**
   The space of biologically relevant VOC is large but is often a result of metabolism, microbiome activity, and the host response [40]; each of these processes generate volatiles through pathways that are increasingly well characterised. Metabolome databases like mVOC [29], HMDB [30], and FooDB [31], combined with constraint-based methods such as flux balance [41] can serve as priors that narrow the search space by predicting which VOCs are likely to be present in a given context. By transferring this mechanistic knowledge into our models, we can expect to reduce the dimensionality.
3. **Recent advances in AI tools could enable us to bridge the gap between analytical chemistry and machine olfaction**
   In 2024, an end-to-end deep learning approach using raw mass spectrometry data achieved up to 0.99 AUC for disease classification without requiring compound identification or traditional biomarker discovery [9]. Transfer learning across instruments and tasks has also been demonstrated [42]. In 2025, Mommers et al. demonstrated molecular structures could be predicted from low resolution GC-MS spectra, which could link GC-MS data to low-dimensional embedding approaches in point 1 [43].
4. **Representations learned on standardised high-resolution data will be of value as deployable sensors mature**
   Miniaturised GC-MS solutions are predicted to be mass market by 2035 [44], increasing the value of signal discovery, transfer learning [42] and representation learning on comparable hardware. New sensing alternatives are emerging such as high-dimensional arrays [45], high chemical specificity [34,46] and online learning sensors [47].

What we expect to fund

We are looking to unlock the "RGB camera + ImageNet" equivalent for digital olfaction. We use the analogy to ImageNet cautiously, as olfaction will not reduce to a small set of primary dimensions the way colour does. The analogy relies on the breakthroughs needed: standardised capture devices + shared data + open benchmarks to create a flywheel that accelerates progress by orders of magnitude.

We intend to fund three workstreams

Workstream A Open Resources Olfactory (ORO)
Workstream B Discovering Olfactory Representations
Workstream C Novel Olfactory Sensing Capability

We will review each of these in turn.

In line with our thesis, we will run workstreams in a phased fashion such that B and C start a fixed period after Workstream A to best utilise the insights from dataset and hardware construction.

Workstream A: Open Resources Olfactory (ORO)
Goal: to build an open, calibrated hardware platform and paired datasets that digital olfaction lacks. Workstream A will draw on the last 40 years of olfactory sensor hardware research, as a first step towards general-purpose olfaction.

We will measure value of the dataset based on its ability to tackle unanswered questions that have been previously raised by the community, including

- The success of high value signal discovery [29].
- The value the dataset brings to sensor design.
- The success of transfer learning between instruments and application datasets.
- The intrinsic dimensionality of olfactory space remains an unanswered scientific question, inflated estimates have been rigorously refuted, but the effective dimensionality is debated [38].

As well as operational metrics to track early signals of success for community engagement, dataset quality, utility, and scale.

We expect this activity will be undertaken by an FRO who will work with creator teams to create the shared resource and solve associated technical challenges. We intend to fund the collection of;

- **Open Source Data:** datasets that concentrate on samples of targeted challenge areas of health, food, and the environment to support building olfactory perception and olfactory signal discovery for competitions and challenges aligned with the programme. We envision datasets will be curated in the lab as well as datasets in the wild; we also see value in using citizen science approaches to collect larger datasets [33].
- **Accessible Hardware:** hardware that can be adopted by researchers alongside associated standards to continue to build on the open source data resource established in the programme.
- **Shared Models:** modelling initiatives that are built on datasets to create synthetic data, learn intermediate representations or pull together 3rd party datasets.
- **Standards and Benchmarks:** a set of standards and benchmarks that support future contributions to the dataset.
- **Sample banking or Biobanking:** where reasonable, systematic collection and cataloguing of physical reference samples such as breath, skin volatiles, food headspace, environmental air paired with rich metadata, so that calibration transfer

and model pre-training can be grounded in shared materials. This has the potential to enable the dataset to evolve beyond current hardware capabilities.

The outcome is the first open, cross-domain dataset of standardised volatile chemical signals paired with application-relevant labels, at a scale sufficient for general-purpose representation learning. The size of the needed dataset is dependent on the dimensionality of olfactory space, with a likely upper bound of approximately one million but with potential representations significantly smaller [28,48]. Market forces alone will not produce this resource, and none currently exist. We will intentionally ensure this resource is accessible beyond the programme and has the governance to set the standard for further data collection in the field.

We are interested in convening this effort, drawing upon expertise from analytical chemistry, olfactory sensing experts, dataset curators, metrologists, sensor manufacturers, and system integration experts. The hardware will encourage contributions for sensor designers keen to open source their designs or for hardware that is readily available for purchase commercially.

The resulting hardware platform and datasets will form the backbone of our research efforts to which other technologies will contribute or benefit from its existence.

Workstream B: Discovering Olfactory Representations
Goal: To discover transferable, low-dimensional representations of volatile chemical space that can be learned from high-resolution data, across connected applications. To achieve our goal, there are important lessons from neural representations in sensory neuroscience, metabolic modeling, flux balance analysis, cross-modal learning, representation learning and AI-driven scientific discovery.

Our basis for confidence is built on recent technical advances and the body of work detailing how biological systems construct flexible low-dimensional representations [49–51]. The Principal Odour Map demonstrates that a graph neural network trained on odour descriptors produced an intermediate representation that predicted human perception of similarity between odours [36]. The Principal Odour Map was trained on single molecules rather than mixtures, but provided early demonstration that a low-dimensional representation was feasible; this work has recently been extended to mixtures of odours [37]. Moreover, the field has been debating the intrinsic dimensionality of this space, with estimates as low as 6, and prohibitively large estimates having been refuted [38]. We note that traditional compound-identification-based biomarker discovery has proven fragile, despite decades of study on lung cancer breath VOCs, no candidate has reached clinical validation, owing to fragmented protocols and methodological heterogeneity across studies [4]. End-to-end learning on raw analytical data has outperformed known markers panels [9].

We will fund teams across academia and industry who are pushing the frontier of how to encode, represent, and learn from the high-dimensional space that is olfaction, building on past work. We are interested in the potential to also incorporate prior knowledge using advancing modelling techniques such as those provided by the Virtual Metabolic Human model [41] and databases like mVOC [29], HMDB [30], FooDB [31], which could predict the VOCs or metabolites likely to be present, further reducing the effective dimensionality of the space.

We expect that teams will work closely with both workstreams A and C. We will measure progress through challenges designed to test cross-domain transfer, intrinsic dimensionality, value of metabolic priors, robustness to confounders, interpretability of learned space and relevance to deployable hardware in Workstream C. We strongly encourage outputs from creators in this workstream to be publicly accessible.

Workstream C: Novel Olfactory Sensing Capability

Goal: To demonstrate deployable sensing capability can capture the discriminative chemical dimensions identified by Workstream B, such that novel solutions can progress to a general-purpose platform of sensors that can address new applications with minimal changes to the core sensor hardware, e.g. via software updates to increase perceptual capabilities.

We anticipate funding teams with deep expertise in deployable sensing technologies. These teams will work closely with Workstream A and B to translate learned representations to drive hardware progress via informative samples, requirements, and testing whether the sensors can resolve the chemical dimensions that carry discriminative information across target domains. We will work with workstream B teams through a series of workshops and draw on past community efforts [28] to define benchmarks testing whether candidate sensors can capture discriminative dimensions from learned representations and generalise across at least two target applications through different models rather than hardware specific modification.

We believe that high-dimensional sensing capabilities hold the greatest potential for general-purpose olfaction. Existing colourimetric sensor arrays have demonstrated that increasing chemical diversity of transduction mechanisms yields an order-of-magnitude increase in independent response dimensions from 2-3 in traditional metal-oxide arrays to over 18 [45], and micro-gas chromatography offers a proven path to miniaturised analytical separation. We will also consider novel approaches that may include, but not limited to the following possibilities:

- biohybrid and synthetic biological receptors
- metal-organic frameworks (MOFs)

- 2D MXenes and heterostructures
- micro-gas chromatography (μGC) on a chip
- neuromorphic olfactory hardware

We welcome proposals with complementary technologies including but not limited to integration of olfaction into multi-modal systems; AI-driven drift compensation and sensor material design; and advanced graph neural networks for mixture decoding.

As shown in the table below, we recognise the importance of improving sensor drift, sensitivity, and selectivity in digital olfaction devices and will favour solutions which are likely to overcome those hurdles in proposals.

Current Capabilities of Digital Olfaction Sensors (adapted from [25])								
	CP	**MOX**	**QCM**	**Nano**	**IMS**	**Optical**	**GC Sensor**	**GC-MS**
Sensitivity	Low	Average	High	Average	High	Average	Average	High
Selectivity	Low	Average	Average	Low	Low	Good	Good	High
Portability	Good	Good	Good	Good	Good	Average	Average	Poor
Cost	Low	Low	Low	Low	Medium	High	Medium	High
Trained personnel	No	No	No	No	No	Yes	Yes	Yes
Throughput	High	High	High	High	Medium	Medium	Medium	Low
Speed	Real-Time	Real-Time	Real-Time	Real-Time	Real-Time /Offline	Real-time/ Offline	Offline	Offline
Chemical insight	No	No	No	No	Yes	Yes	Yes	Yes
Sensor drift	Yes	Yes	Yes	Yes	Minor	Minor	Minor	Minor

Table 1: Technologies include Conducting Polymers (CP), Metal-Oxide (MOS), Quartz Crystal Microbalances (QCM), Nanotube based sensors (Nano), Ion Mobility Spectrometers (IMS), Optical based sensors (Optical), Gas Chromatography-Sensor (GC-Sensor), Gas Chromatography-Mass Spectrometry (GC-MS). Cells coloured in green perform well, orange has acceptable performance, red performs poorly.

We will organise teams around a number of benchmarking milestones and online sync-ups to allow for the exchange of ideas, techniques, and approaches. Teams will be assessed on

a number of metrics, central of which will be the ability to demonstrate transfer learning from the dataset produced in Workstream A; using existing olfactory encodings to better detect new ones.

Measuring Progress: Competitions

Alongside the workstreams, we will create a series of competitions to assess the performance of the creator teams and to engage with the community beyond the scope of the core programme.

Competitions will be designed with input from the creator teams, to push the boundaries of what is possible with an ambitious benchmark to beat. We will fund an independent team to design competitions, set benchmarks and assess progress against the programme metrics. We anticipate running these challenges throughout the programme to assess progress and leverage the data collection workstream.

We also see the competition as an opportunity to engage and excite the wider community engagement with a potential citizen science activity that could enable significantly more in the wild data being acquired which could build on early initiatives like SMELLNET [32] and NY Smells [33], and has been discussed in a recent NSF workshop [28].

What we are still trying to figure out:

We invite feedback and specifically would like to hear input on the following:

- Our selected candidate applications are longitudinal health signals, food spoilage, and food safety. However, we are using this time to select applications within these spaces to optimise for value unlock. We will consider broadening or narrowing the scope on receipt of compelling evidence that suggested domains share a common representation.
- We are seeking input on specific competitions and dataset collection targets. These decisions will be made on ease to collect data at scale, value of application and potential from cross-domain transfer.
- We will work with partners for data collection and data storage beyond the lifetime of the programme. We are starting discussions now for datasets that contain both human-health and non-human health data.
- To be able to learn the language of olfaction, we recognise the size of the dataset is likely to be significant yet vary depending on the application. We are interested in hearing predictions about what size datasets might need to be collected and the reasoning for this?
- Views about relevant standards for open source, publicly accessible, etc for hardware, datasets, and models from the community.
- Views on which instruments we should use to collect the publicly accessible data that will have the most impact in creating a large dataset.

- We would welcome proposed timelines from creators who would be interested in working on representation learning, novel sensor design using data representations, and sensors for high quality lab dataset to help us best assess phasing of workstreams.
- We would welcome proposed timelines from creators who would be interested in developing novel sensor hardware to help us best assess phasing of workstreams.
- How do we manage the open platform such that we achieve maximum gains for the field, whilst providing creators with an opportunity to secure valuable IP? We recognise that enabling IP creation is important to create a sustained economic incentive.

SOURCES

1. van Poelgeest J, Shahbazi Khamas S, Hallawa A, D'Alessandro C, Ferreira R, Maitland-van der Zee AH, et al. Exhaled volatile organic compounds associated with chronic obstructive pulmonary disease exacerbations-a systematic review and validation. J Breath Res. 2025;19. doi:10.1088/1752-7163/adba06

2. Robroeks CM, van Berkel JJ, Jöbsis Q, van Schooten F-J, Dallinga JW, Wouters EF, et al. Exhaled volatile organic compounds predict exacerbations of childhood asthma in a 1-year prospective study. Eur Respir J. 2013;42: 98–106.

3. United Nations Environment Programme. Food Waste Index Report 2024. Think Eat Save: Tracking Progress to Halve Global Food Waste. 2024.

4. Chou H, Godbeer L, Allsworth M, Boyle B, Ball ML. Progress and challenges of developing volatile metabolites from exhaled breath as a biomarker platform. Metabolomics. 2024;20: 72.

5. Anthes E. E-noses Could Make Diseases Something to Sniff at. In: Scientific American [Internet]. 11 Jan 2008 [cited 17 Apr 2026]. Available: https://www.scientificamerican.com/article/electronic-noses-could-make-diseases-something-to-sniff-at/

6. RoboScientific. RoboScientific. In: RoboScientific [Internet]. 24 Nov 2025 [cited 17 Apr 2026]. Available: https://www.roboscientific.com/

7. Expired RFA-TR-21-009: Screening for Conditions by Electronic Nose Technology (SCENT) (U01 Clinical Trial Optional). [cited 17 Apr 2026]. Available: https://grants.nih.gov/grants/guide/rfa-files/RFA-TR-21-009.html

8. Ravia A, Snitz K, Honigstein D, Finkel M, Zirler R, Perl O, et al. A measure of smell enables the creation of olfactory metamers. Nature. 2020;588: 118–123.

9. Deng Y, Yao Y, Wang Y, Yu T, Cai W, Zhou D, et al. An end-to-end deep learning method for mass spectrometry data analysis to reveal disease-specific metabolic profiles. Nat Commun. 2024;15: 7136.

10. Pirrone F, Albertini M. Olfactory detection of cancer by trained sniffer dogs: A systematic review of the literature. J Vet Behav. 2017;19: 105–117.

11. Grandjean D, Sarkis R, Lecoq-Julien C, Benard A, Roger V, Levesque E, et al. Can the detection dog alert on COVID-19 positive persons by sniffing axillary sweat samples? A proof-of-concept study. PLoS One. 2020;15: e0243122.

12. Bauër P, Leemans M, Audureau E, Gilbert C, Armal C, Fromantin I. Remote medical scent detection of cancer and infectious diseases with dogs and rats: A systematic review. Integr Cancer Ther. 2022;21: 15347354221140516.

13. Drabińska N, Flynn C, Ratcliffe N, Belluomo I, Myridakis A, Gould O, et al. A literature survey of all volatiles from healthy human breath and bodily fluids: the human volatilome. J Breath Res. 2021;15: 034001.

14. Totelin L. Smell as sign and cure in ancient medicine. 1st Edition. Smell and the Ancient Senses. 1st Edition. Routledge; 2014. pp. 29–41.

15. Clinical & Evidence-Based Guidelines. [cited 17 Apr 2026]. Available: https://nasemso.org/content.aspx?page_id=22&club_id=157064&module_id=701974

16. Farraia MV, Cavaleiro Rufo J, Paciência I, Mendes F, Delgado L, Moreira A. The electronic nose technology in clinical diagnosis: A systematic review: A systematic review. Porto Biomed J. 2019;4: e42.

17. Sarkar D, Sinclair E, Lim SH, Walton-Doyle C, Jafri K, Milne J, et al. Paper spray ionization ion mobility mass spectrometry of sebum classifies biomarker classes for the diagnosis of Parkinson's disease. JACS Au. 2022;2: 2013–2022.

18. Trivedi DK, Sinclair E, Xu Y, Sarkar D, Walton-Doyle C, Liscio C, et al. Discovery of volatile biomarkers of Parkinson's disease from sebum. ACS Cent Sci. 2019;5: 599–606.

19. Janfaza S, Banan Nojavani M, Khorsand B, Nikkhah M, Zahiri J. Cancer Odor Database (COD): a critical databank for cancer diagnosis research. Database (Oxford). 2017;2017. doi:10.1093/database/bax055

20. Bosch S, Wintjens DSJ, Wicaksono A, Pierik M, Covington JA, de Meij TGJ, et al. Prediction of inflammatory bowel disease course based on fecal scent. Sensors (Basel). 2022;22: 2316.

21. Lourenço C, Turner C. Breath analysis in disease diagnosis: methodological considerations and

applications. Metabolites. 2014;4: 465–498.

22. Pliny WHSJ, Pliny DEE, Pliny, W. H. S. Jones, A. C. Andrews, Pliny HR. Natural History, Volume IV —. In: Harvard University Press [Internet]. [cited 17 Apr 2026]. Available: https://www.hup.harvard.edu/books/9780674994089

23. Darwin C. On the various contrivances by which British and foreign orchids are fertilised by insects, and on the good effects of intercrossing. London: John Murray; 1862.

24. Hernandez-Leyva AJ, Berna AZ, Bui MH, Liu Y, Rosen AL, Lint MA, et al. The gut microbiota shapes the human and murine breath volatilome. Cell Metab. 2026;38: 779–793.e8.

25. Covington JA, Marco S, Persaud KC, Schiffman SS, Nagle HT. Artificial Olfaction in the 21st Century. IEEE Sens J. 2021;21: 12969–12990.

26. Weinberger S. DARPA To Build Cyborg Nose. In: WIRED [Internet]. 30 Nov 2007 [cited 17 Apr 2026]. Available: https://www.wired.com/2007/11/darpa-to-build/

27. Convergence Accelerator Portfolio: Active Tracks. In: NSF - U.S. National Science Foundation [Internet]. [cited 17 Apr 2026]. Available: https://www.nsf.gov/funding/initiatives/convergence-accelerator/portfolio

28. Cleland TA, Covington J, Davis C, Gutierrez-Osuna R, Hanson C, Harris W, et al. Chemical sensing with an olfaction analogue: high-dimensional, bio-inspired sensing and computation. 2022. Available: https://nsf-gov-resources.nsf.gov/2023-03/Chemical%20Sensing%20with%20an%20Olfaction%20Analogue%20High-dimensional%2C%20Bio-inspired%20Sensing%20and%20Computation%20Workshop%20Report_2231512_October%202022_Final.508.pdf

29. Kemmler E, Lemfack MC, Goede A, Gallo K, Toguem SMT, Ahmed W, et al. mVOC 4.0: a database of microbial volatiles. Nucleic Acids Res. 2025;53: D1692–D1696.

30. Wishart DS, Guo A, Oler E, Wang F, Anjum A, Peters H, et al. HMDB 5.0: The Human Metabolome Database for 2022. Nucleic Acids Res. 2022;50: D622–D631.

31. FooDB. [cited 17 Apr 2026]. Available: https://foodb.ca/

32. Feng D, Dai W, Li C, Pernigo A, Wen Y, Liang PP. SmellNet: A large-scale dataset for real-world smell recognition. arXiv [cs.AI]. 2026. doi:10.48550/arXiv.2506.00239

33. Ozguroglu E, Liang J, Liu R, Chiquier M, DeTienne M, Qian WW, et al. New York Smells: A Large Multimodal Dataset for Olfaction. arXiv [cs.CV]. 2025. doi:10.48550/arXiv.2511.20544

34. Breath Biopsy. In: Owlstone Medical - the home of Breath Biopsy® [Internet]. 4 Oct 2023 [cited

17 Apr 2026]. Available: https://www.owlstonemedical.com/

35. Keller A, Gerkin RC, Guan Y, Dhurandhar A, Turu G, Szalai B, et al. Predicting human olfactory perception from chemical features of odor molecules. Science. 2017;355: 820–826.

36. Lee BK, Mayhew EJ, Sanchez-Lengeling B, Wei JN, Qian WW, Little KA, et al. A principal odor map unifies diverse tasks in olfactory perception. Science. 2023;381: 999–1006.

37. Tom G, Ser CT, Rajaonson EM, Lo S, Park HS, Lee BK, et al. From molecules to mixtures: Learning representations of olfactory mixture similarity using inductive biases. arXiv [cs.LG]. 2025. doi:10.48550/arXiv.2501.16271

38. Meister M. On the dimensionality of odor space. Elife. 2015;4: e07865.

39. Jefferis GSXE. Insect olfaction: a map of smell in the brain. Curr Biol. 2005;15: R668–70.

40. Bakkeren E, Piskovsky V, Foster KR. Metabolic ecology of microbiomes: Nutrient competition, host benefits, and community engineering. Cell Host Microbe. 2025;33: 790–807.

41. Noronha A, Modamio J, Jarosz Y, Guerard E, Sompairac N, Preciat G, et al. The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease. Nucleic Acids Res. 2019;47: D614–D624.

42. Seddiki K, Saudemont P, Precioso F, Ogrinc N, Wisztorski M, Salzet M, et al. Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification. Nat Commun. 2020;11: 5595.

43. Mommers J, Barta L, Pietrasik M, Wilbik A. MASSISTANT: A deep learning model for De Novo molecular structure prediction from EI-MS spectra via SELFIES encoding. J Chromatogr A. 2025;1759: 466216.

44. ResearchAndMarkets.com. Trends Shaping the Gas Chromatography Industry, 2025-2035: Portable and Miniaturized GC Systems - ResearchAndMarkets.com. In: Business Wire [Internet]. 21 Apr 2025 [cited 19 Apr 2026]. Available: https://www.businesswire.com/news/home/20250421690152/en/Trends-Shaping-the-Gas-Chromatography-Industry-2025-2035-Portable-and-Miniaturized-GC-Systems--ResearchAndMarkets.com

45. Li Z, Askim JR, Suslick KS. The optoelectronic nose: Colorimetric and fluorometric sensor arrays. Chem Rev. 2019;119: 231–292.

46. McGivern LE, Lim ZH, Yuan Y, Bo Z, Wu G, Bayley H, et al. Targeted high-resolution sensing of volatile organic compounds by covalent nanopore detection. Nat Commun. 2025;16: 9409.

47. Imam N, Cleland TA. Rapid online learning and robust recall in a neuromorphic olfactory circuit. Nat Mach Intell. 2020;2: 181–191.

48. Tripuraneni N, Jin C, Jordan MI. Provable meta-learning of linear representations. arXiv [cs.LG]. 2020. doi:10.48550/arXiv.2002.11684

49. McCalmon H, Cai G, Tsibouris C, Pashakhanloo F, Chung S, Kapoor V, et al. Sparse input representations explain odor discrimination in complex, concentration-varying mixtures. bioRxivorg. bioRxiv; 2026. doi:10.64898/2026.01.27.702074

50. Del Mármol J, Yedlin MA, Ruta V. The structural basis of odorant recognition in insect olfactory receptors. Nature. 2021;597: 126–131.

51. Pashkovski SL, Iurilli G, Brann D, Chicharro D, Drummey K, Franks KM, et al. Structure and flexibility in cortical representations of odour space. Nature. 2020;583: 253–258.

52. Sharma A, Saha BK, Kumar R, Varadwaj PK. OlfactionBase: a repository to explore odors, odorants, olfactory receptors and odorant-receptor interactions. Nucleic Acids Res. 2022;50: D678–D686.

53. Zhao H-Y, Xu S-M, Xie S-N, Ye W-L, Li J, Wang L-H, et al. Atomevo-odor: A database for understanding olfactory receptor-odorant pairs with multi-artificial intelligence methods. Food Chem. 2025;476: 143392.

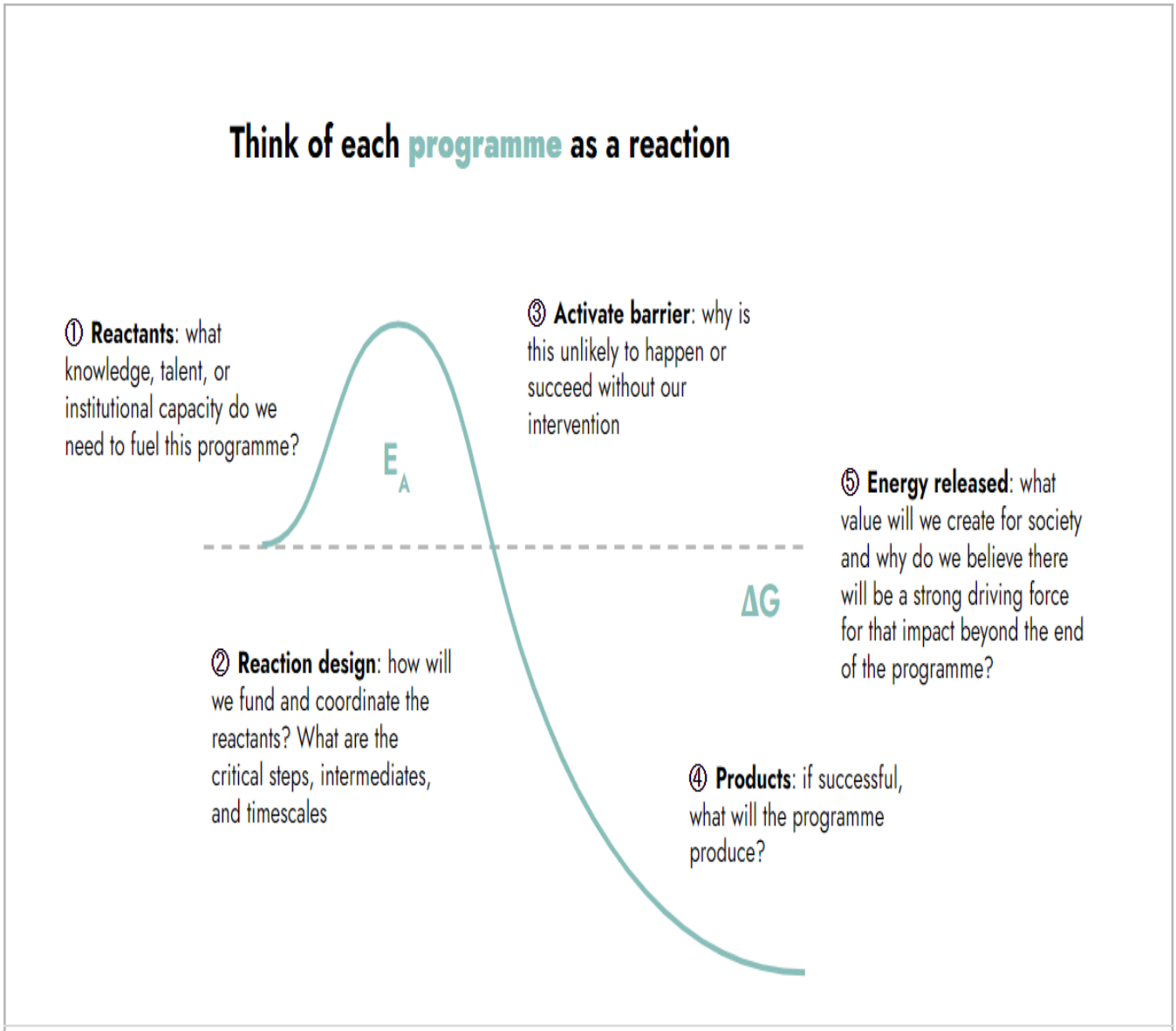
54. Bijland LR, Bomers MK, Smulders YM. Smelling the diagnosis. Neth J Med. 2013;71: 300.

55. Electronic Noses. [cited 17 Apr 2026]. Available: https://warwick.ac.uk/fac/sci/eng/research/grouplist/sensorsanddevices/bsl/e-nose/

56. Cyrano Sciences. [cited 17 Apr 2026]. Available: https://cyranosciences.com/

PROGRAMME THESIS, REACTION DIAGRAM SUMMARY

We can metaphorically think of an ARIA programme as a chemical reaction. We present a simple reaction diagram to summarise the key elements of the imagined programme.



Hypersensory Intelligence : Cracking Digital Olfaction

① Analytical chemistry, deployable olfactory sensing solutions, AI for material sensors design, representation learning, multimodal systems, dataset engineering and curation, metrology, systems integration, domain expertise eg. clinical sciences.

② three workstreams; 1) Open Data, Accessible Hardware and Standards to create the world's largest olfactory dataset; 2) State-of-the-art olfactory sensing capability 3) representation learning to guide bets towards general-purpose hardware and software design. All co-ordinated by a series of programme challenges

③ Market forces overvalue single point solutions and one-off data collection efforts. A unifying force is

needed to organise and standardise digital olfaction.

④ General-purpose olfactory perception capabilities where discovery is enabled by open hardware and software standard alongside a comprehensive dataset all driving commercial solutions.

⑤ Olfactory perception would be a critical part of every sensing package across health, agriculture, environmental sciences, and security.

ENGAGE

Our next step is to launch a funding opportunity derived or adapted from this programme formulation. Click here to register your interest, or to provide feedback that can help improve this programme thesis.

APPENDIX

Related databases

Currently, the data landscape is defined by fragmented repositories: the Cancer Odor Database (COD) for volatile metabolites [19], OlfactionBase [52], HMDB [30], FOODB [31], mVOC [29], and AI-driven mapping platforms like Atomevo-odor [53]. While these resources demonstrate a demand for structured olfactory data, they remain isolated point-solutions. In the age of large, data-driven solutions, we believe now is the time to unify these efforts.

There are also large olfactory perception datasets that have been collected in the wild [32,33].

Introduction to Olfactory Sensing Capability

There are two classes of hardware capabilities to sense volatile chemical signals, those that use lab equipment and those that can be deployed in the field.

In the analytical laboratory, techniques such as GC-MS (Gas Chromatograph Mass Spec), PTR-MS, and SIFT-MS provide detailed chemical profiles, but data acquisition is slow, expensive, and requires highly trained operators. Calibration and reproducibility between instruments remain significant challenges [54], and some volatile compounds readily perceptible to biology are still difficult to capture and quantify analytically.

Solutions that can be deployed outside of the lab that are designed to mimic biological olfaction often use arrays of gas sensors. These sensors have historically struggled with sensor drift, cross-sensitivity, and a lack of standardised data collection protocols [21,54]. Consequently, while we have witnessed the emergence of specific point solutions (from breath alcohol testing to *Helicobacter pylori* detection), we remain far from a general-purpose computational tool for chemical sensing that can match biology.

The lack of standardised and available hardware has led to a market where there are 20 companies that produce 200 machines per year [55]. There are a small number commercially available solutions [56] but the majority of hardware products provide point solutions which are tailored to specific algorithmic applications but no incentives exist for broad purpose olfactory perception to be realised. A general-purpose platform of standardised sensing capability for olfactory perception, stands to open up new applications which cannot be supported by current market dynamics.

Perception vs Discovery of Olfactory Signals

We recognise two ways to address olfactory signals;

1) Olfactory signal discovery: central discovery effort to identify signals that are significant underlying markers and require specific targeted sensing. Previous efforts have struggled with fragmented studies and poor underlying data quality [4,13]. We expect these applications to benefit from centralised signal discovery on standardised high quality lab equipment, which can be informed by contribution in Workstream B such as modelling efforts.
2) Olfactory perception: identifying an intermediate low-dimensional representation or odour map that has been started by neuroscientists and with contributions such as the Principal Odour Map (POM) [36]. There is debate amongst the community about what the true underlying dimensionality is of this space .

This way of dividing the problem mimics biology where pheromone sensing targets specific molecules for a fast response to stimulus versus general olfactory perception which has a more complex neurological structure leveraging the olfactory bulb.