scientific data

Check for updates

OPEN

DATA DESCRIPTOR

MORE-Q, a dataset for molecular olfactorial receptor engineering by quantum mechanics

Li Chen[1], Leonardo Medrano Sandonas[1]✉, Philipp Traber[2], Arezoo Dianat[1], Nina Tverdokhleb[1], Mattan Hurevich[3], Shlomo Yitzchaik[3], Rafael Gutierrez[1], Alexander Croy[2]✉ & Gianaurelio Cuniberti[1,4]✉

We introduce the MORE-Q dataset, a quantum-mechanical (QM) dataset encompassing the structural and electronic data of non-covalent molecular sensors formed by combining 18 mucin-derived olfactorial receptors with 102 body odor volatilome (BOV) molecules. To have a better understanding of their intra- and inter-molecular interactions, we have performed accurate QM calculations in different stages of the sensor design and, accordingly, MORE-Q splits into three subsets: i) MORE-Q-G1: QM data of 18 receptors and 102 BOV molecules, ii) MORE-Q-G2: QM data of 23,838 BOV-receptor configurations, and iii) MORE-Q-G3: QM data of 1,836 BOV-receptor-graphene systems. Each subset involves geometries optimized using GFN2-xTB with D4 dispersion correction and up to 39 physicochemical properties, including global and local properties as well as binding features, all computed at the tightly converged PBE+D3 level of theory. By addressing BOV-receptor-graphene systems from a QM perspective, MORE-Q can serve as a benchmark dataset for state-of-the-art machine learning methods developed to predict binding features. This, in turn, can provide valuable insights for developing the next-generation mucin-derived olfactory receptor sensing devices.

Background & Summary

Introduction. Nowadays, the increasing progress in artificial intelligence (AI) has boosted the development of AI-based technologies for the recognition of objects, faces, voices, and touch[1]. However, a significant gap remains in technologies capable of interpreting and predicting the chemical environment around us. In this sense, tailored electronic noses have recently emerged and are already capable of detecting, for example, volatile organic compounds (VOCs)[2,3]. The detection of VOCs emitted from the human body[4–6], also known as body odor volatilomes (BOVs), can be used as characteristic fingerprints and have significant applications in healthcare[7]. The constituents of BOVs generally indicate the metabolic state of a person and are promising candidates for medical biomarkers in diagnosing a range of diseases[5,8–13], *e.g.* Alzheimer[14,15] and Parkinson[14–17]. A recent report documented that a 'super smeller' could detect and distinguish the BOVs associated with Parkinson's disease, emitted from sebum, from those of normal skin[18] indicating the powerful body odor perception ability enabled by the olfactory system. Given these facts, the demand for fast and robust sensing materials for detecting BOV molecules remains consistently strong, particularly in medical diagnostics.

It is known that odor perception begins in the nose, where tens of thousands of odorants can be detected by the olfactorial receptors, which are composed of a glycoprotein layer (mucin) that covers the epithelium of olfactory and respiratory systems[19–21]. While there have been considerable efforts in investigating single odor (BOV) molecules[22–30] and molecular receptors[31–34], less information is available to accurately describe the physical and chemical interactions in BOV-receptor systems. Characterizing these interactions will deepen our understanding of the biomimetic olfactory system, paving the way for the rational design of receptors tailored for sensing applications[35]. To address this challenge, initial datasets of BOV-receptor systems have been developed[36–40] (see Table 1). For

[1]Institute for Materials Science and Max Bergmann Center for Biomaterials, TUD Dresden University of Technology, 01062, Dresden, Germany. [2]Institute of Physical Chemistry, Friedrich Schiller University Jena, 07737, Jena, Germany. [3]Institute of Chemistry and Center of Nanotechnology, The Hebrew University of Jerusalem, Jerusalem, 91904, Israel. [4]Dresden Center for Computational Materials Science (DCMS), TUD Dresden University of Technology, 01062, Dresden, Germany. ✉e-mail: leonardo.medrano@tu-dresden.de; alexander.croy@uni-jena.de; gianaurelio.cuniberti@tu-dresden.de

Dataset	Receptor source	Total dimer configurations	QM properties	Surface interaction
Mainland *et al.*[38]	Human-cloned	37,303	No	No
OlfactionBase[39]	Human/mouse	875	No	No
M2OR[40]	Mammals	51, 415	No	No
OlfactionDB[37]	Human/mouse	~400	No	No
MORE-Q-G2	Mucin-derived	23,838	Yes	No
MORE-Q-G3	Mucin-derived	1,836	Yes	Yes

Table 1. Main characteristics of recent publicly available olfactory receptor datasets. Note that among these, the MORE-Q dataset uniquely includes quantum-mechanical (QM) properties of the systems under study and is the only dataset that considers the interaction of BOV-receptor systems with a surface.

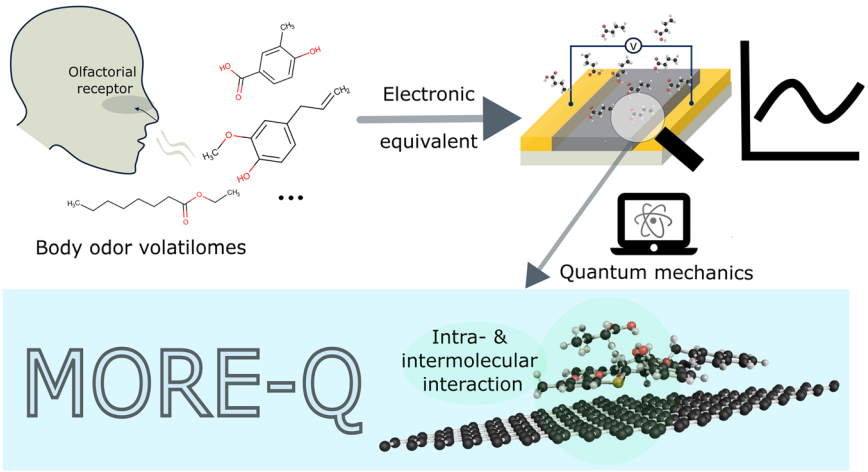

Fig. 1 Graphical representation of the motivation for developing MORE-Q dataset (**M**olecular **O**lfactorial **R**eceptor **E**ngineering by **Q**uantum mechanics). Bio-electronic noses (top right panel) are designed as an electronic equivalent to the olfactory system (top left panel), *e.g.* for sensing body odor volatilomes (BOV). The MORE-Q dataset offers a comprehensive collection of quantum-mechanical properties and structural data that accurately describe intra- and intermolecular interactions in molecular sensors, see lower panel.

instance, Mainland *et al.*[38] provided the *in vitro* response of 73 odorants against a clone library of 511 human olfactorial receptors. Sharma *et al.*[39] also developed the online platform OlfactionBase, which provides chemoinformatic properties (*e.g.* drug-likeness, pharmacokinetic profile, molecular weight, log*P*) and odorant-receptor match information for 875 systems. In a more recent study, Lalis *et al.*[40] developed M2OR database of 51,395 odorant-receptor systems for understanding the molecular mechanisms of olfaction. Numerous similar datasets have been published regarding odorant-receptor interactions and match information[37,41–43]. However, to the best of our knowledge, none of these datasets account for a quantum-mechanical (QM) treatment of physical and chemical interactions in odorant-receptor systems. This is of great relevance because accurately describing intermolecular interactions – such as van der Waals forces, hydrogen bonds, and dipole-dipole interactions – is essential for determining the correct docking configuration and binding features. Indeed, recent efforts have focused on developing QM datasets to understand the structure-property and property-property relationships in both small and large molecular systems[44–49]. Additionally, a few QM datasets for small molecular dimers have been generated, where the relevant property is solely the interaction energy[50–53]. Another open challenge in this field is the accurate investigation of the interaction of odorant-receptor systems on a substrate. Carbon-based materials such as graphene are promising sensing materials due to their high charge mobility and favorable surface-to-volume ratio[54–58]. Hence, it is essential to have a dataset that provides a QM description of both the structural and electronic properties of odorant-receptor systems, as well as their interaction with a sensing material.

To address these challenges, we introduce the MORE-Q dataset, which provides an extensive set of QM properties to accurately investigate the BOV-Receptor systems on a graphene surface, see Fig. 1. We have modeled 18 mucin-derived receptors and their combinations with 102 relevant skin BOV molecules[6]—both systems containing heavy atoms C, N, O, and S. The number of atoms of the receptor molecules ranges from 37 to 102 atoms, while for BOV molecules varies from 7 to 53 atoms. We have followed an exhaustive and systematic procedure to generate the final complex systems (*i.e.* BOV-receptor-graphene system) and compute the QM properties, resulting in the creation of three MORE-Q subsets: i) MORE-Q-G1 contains the QM data of isolated single BOV and receptor molecules, ii) MORE-Q-G2 contains the QM data of diverse configurations of the BOV-receptor systems, and iii) MORE-Q-G3 contains the QM data of dimers deposited on the graphene surface (see Fig. 2). The initial BOV and receptor molecules were optimized using the semi-empirical method GFN2-xTB that considers

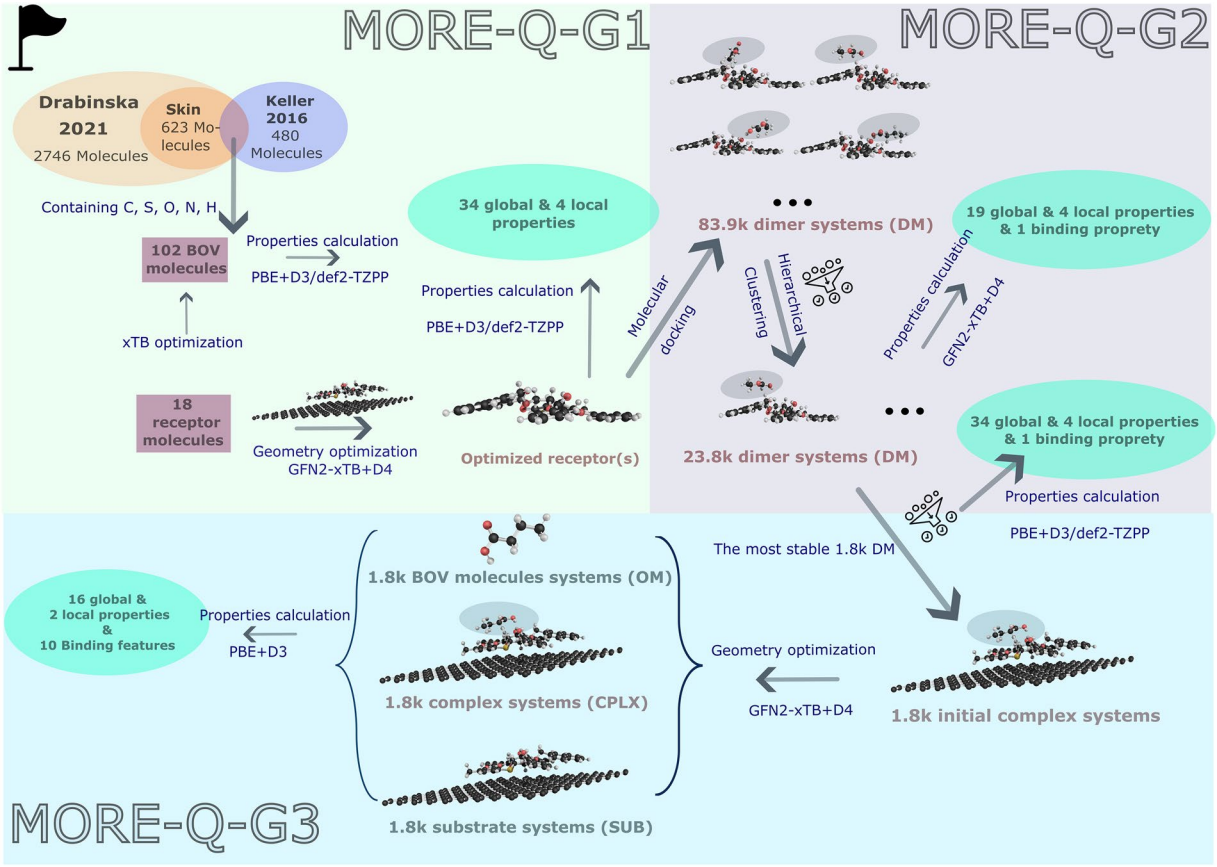

Fig. 2 Schematic description of generation procedure of the MORE-Q dataset. MORE-Q is split into three subsets depending on the generation stage: MORE-Q-G1, MORE-Q-G2, MORE-Q-G3. In MORE-Q-G1, We first established a skin body odor set with 102 molecules from the intersection of two body odor resources such as those presented in Drabinska *et al.*[6] and Keller *et al.*[64]. The structure of isolated BOV molecules and 18 receptor-graphene systems were optimized using GFN2-xTB with D4 dispersion correction. Physicochemical properties were posteriorly computed at the tightly converged PBE+D3 level of theory. For generating MORE-Q-G2, the conformational search of the BOV-receptor systems from MORE-Q-G1 was carried out using the docking program aISS code which employs xTB-IFF[72] force field and a follow-up genetic algorithm, resulting in a total of 83,916 BOV-receptor dimer configurations. Following an RMSD-based hierarchical clustering method, 23,838 non-redundant configurations were selected and their properties were calculated at GFN-xTB+D4 theory level. The most energy-favorable 1,836 configurations were further selected, and their corresponding dimer properties were calculated both at the higher PBE+D3 theory level. In MORE-Q-G3, the selected 1,836 configurations from MORE-Q-G2 were optimized back on the graphene surface, forming the complex BOV-receptor-graphene systems (CPLX). The substrate (SUB) and BOV molecule (OM) systems were constructed by removing the BOV molecules and receptor+graphene, respectively. Next, PBE+D3 level simulations were conducted on the geometries of 1,836 CPLX, SUB, and OM systems, obtaining global and local properties, as well as the binding features of these systems. See "Methods" for more details.

D4 dispersion correction[59]. Then, the configurations of the 1,836 BOV-receptor systems were screened by molecular docking using the automated Interaction Site Screening (`aISS`)[60] submodule of the xTB code, applying the same level of theory as used in the geometry optimizations. Here, the hierarchical clustering method was used to filter similar geometries for each BOV-receptor combination, resulting in 23,838 dimer configurations. To generate the complex systems, we have considered the most energy-favorable dimer configuration for each unique combination (ranked by the interaction energy E_{int}) and, then, deposited it on a graphene layer. Finally, each MORE-Q subset includes up to 39 physicochemical properties, encompassing global (molecular) and local (atom-in-a-molecule) properties, as well as binding features. All these properties were computed using the tightly converged PBE+D3 level of theory. As such, the MORE-Q dataset provides a comprehensive set of QM structural and property data for BOV-receptor and BOV-receptor-graphene systems, which can further enhance our understanding and the prediction of the performance of molecular sensors for digital olfaction.

Key advancements. The MORE-Q dataset series aims to provide accurate QM data to gain insights into the interaction between BOV molecules and molecular receptors, as well as the effects of substrate deposition on binding features. To achieve this objective, we have ensured that MORE-Q includes the following attributes:

- The MORE-Q dataset comprises the QM structural and property data of 18 newly synthesized mucin-derived receptors[32,61,62] and 102 skin molecules, carefully selected from an extensive pool of 2, 746 BOV molecules, representing a large swath of the chemical space of BOV molecules.
- We have exhaustively explored the potential energy surface (PES) of BOV-receptor systems by using the automated interaction site screening method in xTB code[60], at the GFN2-xTB level of theory[59] with D4 dispersion correction[63].This procedure allowed us to determine the 1,836 most energetic-favorable BOV-receptor dimer configurations.
- The MORE-Q dataset provides extensive sets of QM global and local properties (up to 39) for single BOV/receptor molecules (MORE-Q-G1), molecular dimers (MORE-Q-G2), and complex systems (MORE-Q-G3), offering more comprehensive data compared to all other aforementioned BOV-receptor interaction datasets. These properties can be utilized in machine learning (ML) methods (*e.g.* as QM descriptors) to uncover and elucidate structure-property and property-property relationships between the building blocks, as well as the entire molecular sensor.
- The MORE-Q-G3 dataset involves complex electronic structure calculations, such as the determination of the work function (WF) ϕ, which can be related to the sensing response of molecular devices. This property offers a more realistic validation of the ML models trained on our QM data, intending to design novel molecular sensors.

Methods

BOV and receptor molecules. The selection of the initial set of BOV molecules dataset was based on two main works[6,64]. In the first one, Drabińska *et al.*[6] reports a meta-analysis of the available literature on chemical substances that have been documented in the human body. Here, 2,752 molecules are categorized into feces, urine, breath, skin, milk, blood, saliva, and semen. Since a central aspect of digital olfaction is the perception of odor molecules, we have also examined the work done by Keller *et al.*[64], and consider 480 BOV molecules with available perceptions from their dataset. Then, by intersecting both datasets, we ended up with 102 skin-related molecules that include the heavy elements C, S, O, and N (see Fig. 2). Their isomeric SMILE strings and the corresponding initial geometries were extracted from the large public molecular database PubChem[65]. The molecular size of the BOV molecules varies from 7 to 53 atoms. The two-dimensional chemical representation of these molecules has been plotted in Fig. S1 of the Supporting Information (SI). Furthermore, we have modeled newly-synthesised 18 bio-inspired (mucin-derived) receptors[32,61,62] that are composed of glycan modified by aromatic decoration for surface adhesion. D-galactose, one of the most common glycans in the extracellular matrix, was used as a scaffold for aromatic decorated monosaccharide receptor's library, which is obtained by a multistep chemical synthesis[62] of monosaccharides. The synthetic approach provides the ability to install specific groups of various natures on the monosaccharide thereby enabling tuning the receptor's affinity toward odorants. Using this ability we control the rigidity, hydrophobicity, and polarizability of glycan-based receptors. The molecular size of the receptors ranges from 37 to 102 atoms, including the heavy atoms C, N, O, and S. A detailed description and full characterization can be found in Refs. [32,62]. We show the two-dimensional chemical structures of the 18 receptors in Fig. S2 of the SI.

MORE-Q-G1 dataset generation: properties of monomers. The MORE-Q-G1 dataset contains the quantum-mechanical (QM) properties of the optimized structures of 102 BOV molecules and 18 molecular receptors. The structures of each BOV molecule were first optimized using the semi-empirical method GFN2-xTB that considers D4 dispersion correction as it is implemented in the xTB packages (version 6.6.0)[59]. A stringent convergence criterion for energies and gradient norms was set to $5 \times 10^{-8} E_h$ and $5 \times 10^{-5}\ E_h \cdot a_0^{-1}$, respectively. For the molecular receptors, we deposited them directly on a 10×10 graphene layer containing 200 C atoms with the fixed vacuum layer $z = 50.68$ Å in a periodic simulation box. The geometry optimization of the receptor-graphene systems (referred to as the 'SUB' system in other sections) was conducted using DFTB+ package[66,67] and considering the GFN2-xTB Hamiltonian with D4 dispersion correction. The thresholds for SCC convergence and the maximal atomic force were set to $1 \cdot 10^{-5}$ and $1 \cdot 10^{-4}\ E_h \cdot a_0^{-1}$, respectively. The Fermi smearing in the optimization was set to 300 K and simulation was conducted at Gamma point. During the optimization, lattice vector angles and slab thickness were fixed. For the initial adsorption of each receptor on graphene, configurations were set to have maximal $\pi - \pi$ stacking between the pyrene rings of the receptors and the graphene layer. Notice that, instead of optimizing the receptors as an isolated system, we directly optimized them within the periodic graphene system to avoid self $\pi - \pi$ stacking in the isolated state, which would impede stable adsorption on the graphene surface.

Calculation of physicochemical properties. We have computed 39 QM global and local properties of the optimized structures of BOV and receptor molecules, see Table 2. To do this, single-point calculations were conducted employing density-functional theory (DFT) at the PBE level with def2-TZVPP basis set and D3 dispersion correction, as implemented in the ORCA software (version 5.0.3)[68]. Energy components, orbital energies, C_6 dispersion coefficient, atomic forces, dipole moment, and quadrupole moment were extracted from the output files of the self-consistency (SCF) calculations. The molecular isotropic polarizability and the polarizability tensor were analytically calculated through the coupled-perturbed SCF equations (CP-SCF). The radius of gyration was calculated per each structure by $R_g = \frac{\Sigma m_i \cdot r_i^2}{\Sigma m_i}$, where m_i and r_i are the i^{th} atom mass and the corresponding distance to the molecular center of mass, respectively. The atomic charges were computed by performing Mulliken[69], Loewdin[70] and Mayer[71] population analysis.

#	Property	Symbol	Unit	Dimension	Type	HDF5 keys
1	Atomic numbers	—	—	N	A	'atNUM'
2	Atomic positions	—	Å	3N	A	'atXYZ'
3	Total PBE+D3 energy	E_{tot}	eV	1	M	'ePBE+D3'
4	Nuclear repulsion energy	E_{nuc}	eV	1	M	'eNUC'
5	Electronic repulsion energy	E_{ele}	eV	1	M	'eELE'
6	One electron energy	E_{1e}	eV	1	M	'e1E'
7	Two electron energy	E_{2e}	eV	1	M	'e2E'
8	Virial potential energy	E_{pe}	eV	1	M	'ePE'
9	Virial kinetic energy	E_{ke}	eV	1	M	'eKE'
10	Exchange energy	E_x	eV	1	M	'eX'
11	Correlation energy	E_c	eV	1	M	'eC'
12	Exchange-correlation energy	E_{xc}	eV	1	M	'eXC'
13	Total D3 energy	E_{D3}	eV	1	M	'eD3'
14	Dispersion E6 energy	E_6	eV	1	M	'eE6'
15	Dispersion E8 energy	E_8	eV	1	M	'eE8'
16	HOMO energy	E_{HOMO}	eV	1	M	'eH'
17	LUMO energy	E_{LUMO}	eV	1	M	'eL'
18	HOMO-LUMO gap	E_{GAP}	eV	1	M	'HLgap'
19	Orbital energies	E_{oe}	eV	*	M	'eORB'
20	Isotropic molecular C_6 coefficient	C_6	$E_h \cdot a_0^6$	1	M	'mC6'
21	Electronic dipole moment	μ_{elc}	D	3	M	'vEDIP'
22	Nuclear dipole moment	μ_{nuc}	D	3	M	'vNDIP'
23	Total dipole moment	μ	D	3	M	'vDIP'
24	Scalar total dipole moment	μ_s	D	1	M	'DIP'
25	Rotational spectrum constant	B	MHz	3	M	'vRS'
26	Rotational dipole moment	μ_B	d	3	M	'vRSDIP'
27	Nuclear quadrupole moment tensor	Q_{nuc}	$e \cdot a_0^2$	6	M	'NQP'
28	Electronic quadrupole moment tensor	Q_{ele}	$e \cdot a_0^2$	6	M	'EQP'
29	Total quadrupole moment tensor	Q	Buckingham	6	M	'TQP'
30	Isotropic molecular quadrupole	Q_s	Buckingham	1	M	'mQP'
31	Molecular polarizabillity tensor	α	a_0^3	6	M	'mTPOL'
32	Molecular isotropic polarizability	α_s	a_0^3	1	M	'mPOL'
33	Radius of gyration	R_g	Å	1	M	'RG'
34	Inertia moment tensor	I_{TS}	amu ·Å^2	6	M	'IM'
35	Mulliken atomic charge	q_{mu}	e	N	A	'muCHG'
36	Loewdin atomic charge	q_{lo}	e	N	A	'loCHG'
37	Mayer atomic charge	q_{ma}	e	N	A	'maCHG'
38	Atomic forces	F_{at}	eV/Å	3N	A	'vF'
39	Atomisation energy	E_{at}	eV	1	M	'eAT'

Table 2. List of physicochemical properties of BOV and molecular receptors contained in MORE-Q-G1 subset. Each property presents a name, symbol, unit, dimension, type, and corresponding key in the HDF5 file. Property types are categorized into atomic (A) and molecular (M). $E_h$and a_0 refer to the atomic unit of Hatree and Bohr radius. * The number of orbital energies varies for each molecule.

MORE-Q-G2 dataset generation: properties of the molecular dimers. The MORE-Q-G2 dataset contains the QM properties of the optimized structures of the 23,838 BOV-receptor systems (referred to as the dimer system). The configurations of the initial 1, 836 BOV-receptor systems (combination of 18 molecular receptors with 102 BOV molecules from MORE-Q-G1) were screened by molecular docking using the automated Interaction Site Screening (`aISS`)[60] submodule of the xTB packages (version 6.6.0), where the GFN2-xTB parameterization and D4 dispersion correction were implemented. The `aISS` module prescreens potential docking sites (pockets, stack, and angular search) on the receptor, followed by a genetic optimization for stack and angular search, where the intermediate binding energies are evaluated using the xTB-IFF force field[72]. In the end, the updated dimer structures were optimized again by the GFN2-xTB method. To keep the same structural conformation of the receptor adsorbed on the graphene layer, we have fixed the geometry of the receptor during the docking process. Then, 100 configurations were generated per molecular dimer. We subsequently sent the dimer configurations back to the graphene surface and excluded those configurations in which any atom of the BOV molecule was located between the receptor and the graphene layer, resulting in a subset of 83,916 dimer configurations. It is worth mentioning that the atomic coordinates of the receptor were mapped exactly to the previous

#	Property	Symbol	Unit	Dimension	Type	HDF5 keys
1	Atomic number	—	—	N	A	'atNUM'
2	Atomic positions	—	Å	3N	A	'atXYZ'
3	Total GFN2-xTB+D4 energy	E_{tot}	eV	1	M	'eXTB+D4'
4	Repulsion energy	E_{rep}	eV	1	M	'eREP'
5	SCC total energy	E_{scc}	eV	1	M	'eSCC'
6	Isotropic electrostatic energy	E_{iel}	eV	1	M	'eIE'
7	Anisotropic electrostatic energy	E_{ael}	eV	1	M	'eAE'
8	Anisotropic exchange-correlation energy	E_{axc}	eV	1	M	'eAXC'
9	D4 dispersion energy	E_{D4}	eV	1	M	'eD4'
10	HOMO energy	E_{HOMO}	eV	1	M	'eH'
11	LUMO energy	E_{LUMO}	eV	1	M	'eL'
12	HOMO-LUMO gap	E_{GAP}	eV	1	M	'HLgap'
13	Orbital energies	E_{oe}	eV	*	M	'eORB'
14	Atomisation energy	E_{at}	eV	1	M	'eAT'
15	Atomic coordination number	N_{ac}	—	N	A	'ACN'
16	Atomic Mulliken charge	q_{mu}	e	N	A	'muCHG'
17	Atomic C_6 dispersion coefficient	$C_{6,\text{at}}$	$E_{\text{h}} \cdot a_0^6$	N	A	'atC6'
18	Atomic polarizability	α_{at}	a_0^3	N	A	'atPOL'
19	Isotropic C_6 dispersion coefficient	C_6	$E_{\text{h}} \cdot a_0^6$	1	M	'mC6'
20	Isotropic C_8 dispersion coefficient	C_8	$E_{\text{h}} \cdot a_0^8$	1	M	'mC8'
21	Isotropic molecular polarizability	α_{s}	a_0^3	1	M	'mPOL'
22	Dipole moment	μ	$e \cdot a_0$	3	M	'vDIP'
23	Scalar total dipole moment	μ_{s}	D	1	M	'DIP'
24	Molecular quadrupole tensor	QP	$e \cdot a_0^2$	6	M	'QP'
25	Binding energy	E_{int}	eV	1	M,BD	'eBIND'

Table 3. List of physicochemical properties of molecular dimers at GFN2-xTB+D4 theory level contained in MORE-Q-G2 subset. Each property presents a name, symbol, unit, dimension, type, and corresponding key in the HDF5 file. Property types are categorized into atomic (A) and molecular (M). BD stands for the binding feature. E_{h} and a_0 refer to the atomic unit of Hatree and Bohr radius. For the most stable dimer conformation, we have also computed the same properties as listed in Table 2 at PBE+D3 level of theory, including the binding energy. *The number of orbital energies varies for each molecule.

configuration on graphene from the MORE-Q-G1 dataset. As a final step, the hierarchical clustering method was used to filter similar geometries for each BOV-receptor combination, reducing the number of configurations up to 23,838. This step was carried out by computing the root-mean-square deviation (RMSD) among molecular structures. The detailed clustering process is discussed in Sec. 2 of the SI.

Calculation of physicochemical properties. We have first computed 24 QM global and local properties of the optimized structures of 23,838 BOV-receptor systems, see Table 3. These properties were obtained from the output files of a follow-up single-point calculation using GFN2-xTB, which considers D4 dispersion correction. The SCC convergence for these calculations was set to $1 \cdot 10^{-6}\, E_{\text{h}}$. To name a few properties, we have energy components, orbital energies, C_6 and C_8 dispersion coefficients, atomic polarizabilities, dipole moment, quadrupole moment, and binding energy. Moreover, we have selected the most energy-favorable configuration for each dimer (ranked by the binding energy E_{int}) and computed QM properties at the PBE+D3 level with def2-TZVPP basis set, as was previously done for the generation of MORE-Q-G1. Accordingly, MORE-Q-G2 also contains the 39 QM global and local properties listed in Table 2 and E_{int} for 1,836 dimers at PBE+D3 level.

MORE-Q-G3 dataset generation: properties of the complex systems. The MORE-Q-G3 dataset contains the QM properties of the optimized structures of 1, 836 BOV-receptor-graphene systems (referred to as the complex (CPLX) system). To generate MORE-Q-G3, we have considered the most energy-favorable dimer configuration for each dimer (ranked by E_{int} from MORE-Q-G2) and, then, mapped it back to the graphene layer. Next, the CPLX systems underwent geometry optimization using the DFTB+ package, employing the GFN2-xTB Hamiltonian with D4 dispersion correction for the SUB system. We here chose to fix the atomic positions in the graphene layer, as the adsorption of BOV molecules will not significantly affect them. The resulting 1,836 CPLX systems were then split into SUB systems and BOV molecules (OM) to compute binding features.

Calculation of physicochemical properties. We have first computed 20 QM global and local properties of the optimized structures of 1,836 CPLX systems, 1,836 SUB systems, and 1,836 OM systems, see the top panel in Table 4. In doing so, single-point calculations of these systems were conducted at tightly converged PBE+D3 theory level by Vienna ab initio simulation package[73,74] (VASP, version 6.3.1). The energy cutoff for

#	Property	Symbol	Unit	Dimension	Type	HDF5 keys
1	Atomic number	—	—	N	A,S	'atNUM'
2	Atomic coordinates	—	Å	3N	A,S	'atXYZ'
3	Total PBE+D3 energy	E_{tot}	eV	1	G,S	'ePBE+D3'
4	Fermi energy	E_F	eV	1	G,S	'eFE'
5	E6 dispersion energy	E_6	eV	1	G,S	'eE6'
6	E8 dispersion energy	E_8	eV	1	G,S	'eE8'
7	Total dispersion energy	E_{D3}	eV	1	G,S	'eD3'
8	Valence band maximum	E_{vbm}	eV	1	G,S	'eVBM'
9	Conduction band minimum	E_{cbm}	eV	1	G,S	'eCBM'
10	VBM-CBM gap	E_{gap}	eV	1	G,S	'eGAP'
11	Band energies	E_{be}	eV	*	G,S	'eBE'
12	Work function	ϕ	eV	1	G,S	'WF'
13	Planar-averaged potential z distance	z	Å	*	G,S	'zEPOI'
14	Planar-averaged potential	P_{avg}	eV	*	G,S	'eEPOI'
15	Cell parameters	l	Å	9	G,S	'CELL'
16	Cell stress tensor	σ	kB	6	G,S	'stCELL'
17	External cell pressure	P_{cl}	kB	1	G,S	'pCELL'
18	Atomic forces	F_{at}	eV/Å	3N	A,S	'vF'
19	Total drift	F_{df}	eV/Å	3	A,S	'vDF'
20	Bader atomic charge	q	e	N	A,S	'baCHG'
1	Adsorption energy	E_{ads}	eV	1	G,BD	'eADS'
2	Graphene Bader charge change	ΔQ_{GR}	e	1	G,BD	'GbaDCHG'
3	Receptor Bader charge change	ΔQ_{rec}	e	1	G,BD	'RbaDCHG'
4	BOV molecule charge change by Bader analysis	ΔQ_{om}	e	1	G,BD	'ObaDCHG'
5	Work function change	$\Delta\phi$	eV	1	G,BD	'DWF'
6	Dispersion energy change	ΔE_{D3}	eV	1	G,BD	'DD3'
7	Electronic gap change	ΔE_{gap}	eV	1	G,BD	'DGAP'
8	Bader atomic charge change	Δ_q	e	*	A,BD	'baDCHG'

Table 4. List of physicochemical properties of molecular systems at PBE+D3 theory level contained in MORE-Q-G3 subset. Each property presents a name, symbol, unit, dimension, type, and corresponding key in the HDF5 file. Property types are categorized into atomic (A) and global (G). S and BD stand for a single system (*e.g.* CPLX, SUB, and OM) and for the binding feature, respectively. E_h and a_0 refer to the atomic unit of Hatree and Bohr radius. *The dimension of these properties varies for each molecule.

the plane-wave basis set and the SCF convergence threshold were set to 600 and $1 \cdot 10^{-5}$ eV, respectively. And all simulations were conducted at Gamma point. The dipole correction along the slab direction (50.68 Å) was switched on to obtain flat electrostatic potential in the slab. Energy components, orbital energies, atomic forces, stress tensor, and electrostatic potential were extracted from the OUTCAR output file. Bader atomic charges q were obtained by postprocessing the information obtained from Bader charge analysis[75].

The work function of the CPLX and SUB systems were calculated as follows:

$$\phi = E_V - E_F, \tag{1}$$

where E_F is the Fermi level and E_V is the vacuum energy. And the work function change is defined as the work function difference after and before the BOV adsorption *i.e.* CPLX and SUB systems:

$$\Delta\phi = \phi_{CPLX} - \phi_{SUB}. \tag{2}$$

E_V is obtained by analyzing the flattened region of the electrostatic potential $P(z)$ along the slab direction. $P(z)$ is computed by the following equation:

$$P(z) = \int n(z)dz, \tag{3}$$

where the planar averaged charge density $n(z)$ is defined as:

$$n(z) = 1/A \iint n(x, y, z)dxdy \tag{4}$$

and the A denotes the surface area of the cell.

To investigate the sensitivity and selectivity of the receptors, we have also calculated 8 binding features for these systems, see the bottom panel in Table 4. For example, the adsorption energy E_{ads}, which is defined as the interaction strength between OM and SUB systems, was computed by evaluating:

$$E_{ads} = E_{CPLX} - E_{SUB} - E_{OM}, \quad (5)$$

where E_{CPLX}, E_{SUB}, and E_{OM} are the total energy of each system obtained by VASP. The atomic charge change Δq is obtained by the difference between the Bader atomic charge q of the same atom in CPLX and SUB (OM) systems. By summing up the atomic charge changes of the atoms for individual components, we can obtain the charge change for the receptor (ΔQ_{rec}), BOV molecule (ΔQ_{om}), and graphene substrate (ΔQ_{GR}) upon adsorption of BOV molecules. The other binding features were computed similarly, taking into account the values obtained for the CPLX and SUB systems.

Interconnection between the MORE-Q subsets. Here, we summarize the interactions among the three MORE-Q subsets in the following points:

- The MORE-Q-G1 subset contains QM property data for 102 BOV molecules and 18 molecular receptors. Among the 39 molecular and atomic properties, we computed the D3 energy, dipole moment, polarizability, and Mulliken charges (see the property list in Table 2).
- The MORE-Q-G2 subset is built on the geometries from MORE-Q-G1 via the search for molecular docking conformations using BOV molecules and receptors. Accordingly, MORE-Q-G2 contains QM property data for 23,838 dimer conformations at the GFN2-xTB+D4 level and for 1,836 dimers with the lowest binding energies at the PBE+D3 level (see the property list in Tables 2 and 3).
- The MORE-Q-G3 subset is constructed by depositing 1,836 selected dimers from MORE-Q-G2 onto a graphene surface. Consequently, MORE-Q-G3 includes QM property data at the PBE+D3 level for both the CPLX and SUB systems, as well as binding features that account for property changes in single systems induced by BOV molecule adsorption (see the property list in Table 4).

Data Records

The MORE-Q datasets are available in three HDF5 files in the ZENODO.ORG data repository[76]. Indeed, one can find there the files `MORE-Q-G1.hdf5`, `MORE-Q-G2.hdf5`, and `MORE-Q-G3.hdf5` corresponding to the three MORE-Q datasets described in this work. We also provide a README file with technical usage details and examples of how to extract data from the HDF5 files and, then, convert it to Python pandas Dataframes for further analysis (see `createDF.py` file)

HDF5 file format. *File sutrcture.* Independent of the MORE-Q subset, the information for each molecular structure is stored in a Python dictionary (`dict`) type containing all relevant properties and recorded in *groups* in HDF5 file format. The HDF5 file architecture of the MORE-Q subsets is depicted in Fig. 3.

- For `MORE-Q-G1.hdf5` file, containing QM properties of BOV molecules (om) and molecular receptors (rec), `nn_om` and `mm_rec` are allocated to the main *groups* keys, where `nn` and `mm` range from 1 to 102 and from 1 to 18, respectively. HDF5 keys to access the atomic numbers, atomic positions (coordinates), and physicochemical properties in each dictionary are provided in Table 3.
- For `MORE-Q-G2.hdf5` file, containing QM properties of the molecular dimers at different levels pf theory, `XTB` and `ORCA` become the main *groups* keys. Under both of them, `mm_rec` is created as the subgroup. Then, a `nn_om` subgroup is created per `mm_rec` subgroup, where `nn_om` and `mm_rec` represent the combination of BOV molecules (om) and molecular receptors (rec) that compose a given molecular dimer. Following this, a third-level subgroup, `ll_dm`, is created to indicate the dimer configurations within each `nn_om` subgroup, where `ll` denotes the order of the dimers. Under the `ll_dm` subgroup, `DM` (dimer), `OM` (isolated BOV molecule), and `REC` (receptor) subgroups were created to store the QM properties listed in Tables 3 and 2 for each system. The binding energy values are saved in the `DM` subgroup.
- For `MORE-Q-G3.hdf5` file, containing QM properties of the complex systems (*i.e.* BOV-receptor-graphene), substrate systems (*i.e.* receptor-graphene), BOV molecules, and the binding features, the main *groups* keys `CPLX`, `SUB`, `OM` and `BD` are constructed. Alike the structure of `MORE-Q-G2.hdf5` file, we have first created the `mm_rec` subgroup for each main group. Then, the subgroup, `nn_om`, is created per `nn_om` subgroup, representing the BOV molecular and molecular receptor combination. Followed by `ll_dm`, the dimer configuration order is indicated. QM properties listed in Table 4 are then stored per subgroup.

Property format. MORE-Q HDF5 files contain various types of data derived from the features of each property. The atomic numbers are stored as a list of strings, with each string representing the atomic number of a corresponding atom in the molecule. The atomic coordinates and forces are saved as a list of 3N vectors, where each vector contains the x, y, and z components. Orbital (band) energies are saved as a vector array, containing 10 energies below HOMO (VBM) level and 10 energies above the LUMO (CBM) level. Vectorial and tensorial properties are also saved as a vector array, with the order maintained as `x, y, z` for vectorial properties and `xx, yy, zz, xy, yz, zx` for tensorial properties. For all atomic properties, the order of the vector array is identical to the atomic coordinates vector. For the MORE-Q-G2 and MORE-Q-G3 subsets, the order of the

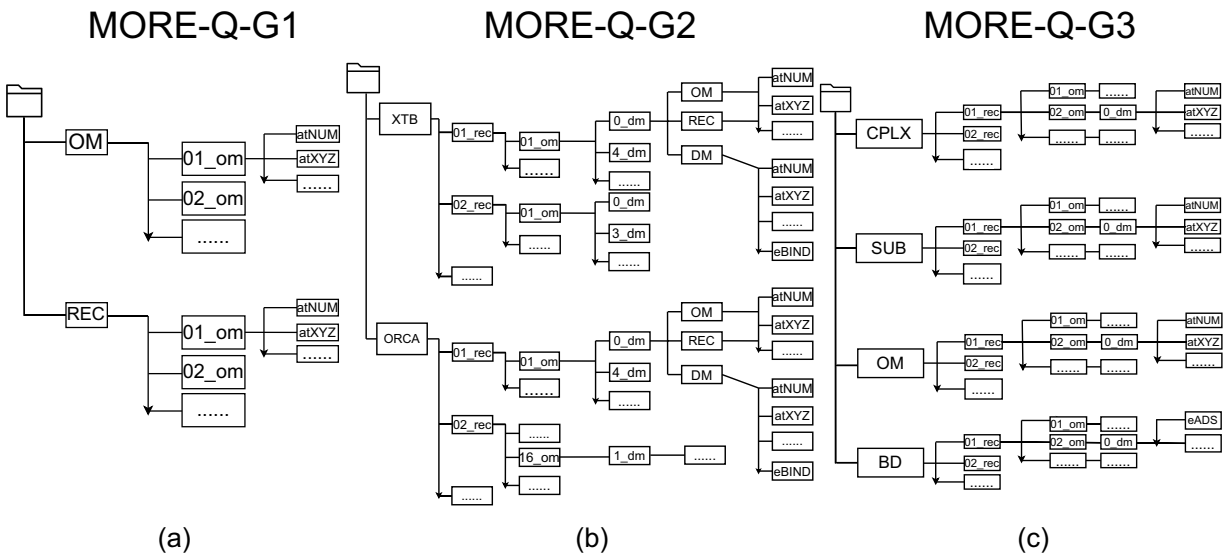

Fig. 3 Architecture of the HDF5 files corresponding to (**a**) MORE-Q-G1, (**b**) MORE-Q-G2, and (**c**) MORE-Q-G3 subsets.

atom type follows receptor → BOV and graphene → receptor → BOV, respectively. This order is also applicable to every atomic property. The other properties are single values that could be directly called by the corresponding HDF5 keys, see Tables 2–4.

Technical Validation

Unlike other datasets in the field of digital olfaction, the MORE-Q datasets contain an extensive set of quantum-mechanical (QM) properties for the building blocks of graphene-based molecular sensors, as well as binding features among the sensor components (see Fig. 2). Indeed, structural properties were obtained using the semi-empirical GFN2-xTB method with D4 dispersion correction[63], which is known to generate correct geometries at an efficient computational cost. While the energetic, atomic forces and other property calculations were performed at the more accurate level of theory such as PBE with D3 dispersion correction[77], as implemented in the VASP code. By doing this, we have guaranteed both sufficiently accurate QM properties and feasible computational time consumption.

Before constructing our complex CPLX system (*i.e.* BOV-receptor-graphene system), we have determined the optimal configuration for the molecular receptor adsorbed on the graphene layer. In this regard, it is known that $\pi - \pi$ stacking interaction is the main functionalization mechanism of the mucin-derived receptor. Accordingly, the receptor configuration with the maximal pyrene rings interacting with graphene will be the most stable receptor-graphene system (referred to as the SUB system). Based on this concept, we initially deposited each receptor on graphene in up to four configurations with diverse orientations, which were subsequently optimized using GFN2-xTB methods with D4 correction and tight settings of convergence. Then, the most stable configurations with the lowest adsorption energy were taken as the backbone structures, which were used to further generate the MORE-Q datasets, as shown in Fig. 2. Note that a more exhaustive evaluation of the conformational space of the molecular receptors may yield different results. However, the main focus of the current work is to define the pathways for understanding and predicting QM-based structure-property and property-property relationships in potential molecular sensors for digital olfaction. Thus, the selection of these configurations represents the most likely configurations and, therefore, ensures the baseline quality of the dataset.

Next, the selected receptor configurations were combined with BOV molecules to form the input geometries for the `aISS` code[60] to determine the docking sites. This method has been successfully used to determine the explicitly solvated structures of peptides and macrocycles from the MPCONF196 dataset[78]. After performing the docking procedure, hierarchical clustering[79] based on root-mean-squared deviation (RMSD) and energetics was carried out to select non-redundant configurations, resulting in a total of 23,838 molecular dimers (see more details for clustering in Fig. S3 of the SI). From this subset, we have selected the most energy-favorable configuration per molecular dimer (ranked by the binding energy E_{int}) for further examination. Indeed, the principal component analysis (PCA) is conducted on the 23,838 and the selected 1,836 configurations using the global properties stored in MORE-Q-G2 dataset. Fig. 4(a) shows the PCA space for both molecular sets, where one can see that the selected 1, 836 configurations span the entire region covered by the 23,838 dimers. This indicates that the reduced set is a representative sample of the dimer configurations. As shown in Fig. 4(b), even though we sampled only configurations with the lowest E_{int}, the coverage of the initial E_{int} values is considerable when considering the reduced molecular set. Here, configurations with $|E_{\text{int}}|$ smaller than 0.25 eV were filtered out after the energetic selection, as these meta-stable dimers exhibit weak non-covalent interactions and are unlikely to occur during the docking process — another compelling evidence that the reduced set is a representative

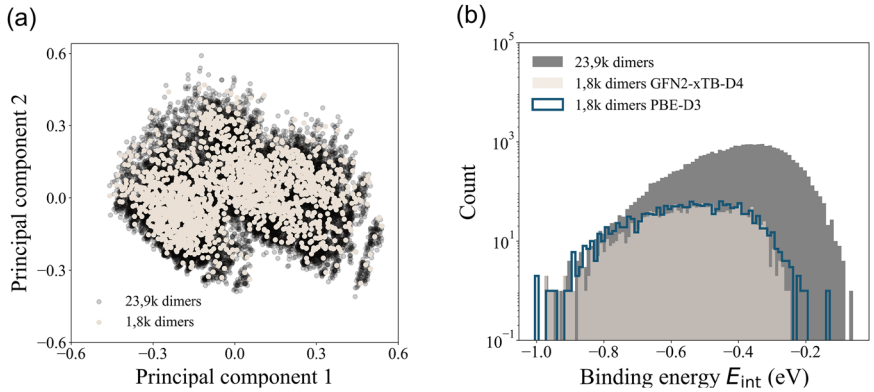

Fig. 4 Analysis of the conformational and energetic space for molecular dimers in MORE-Q-G2 subset. (**a**) Two-component principal component analysis based on the global properties for the 23,838 (black circles) dimer conformations stored in MORE-Q-G2. For comparison, we show the values corresponding to the 1,836 most energetically stable dimer conformation (white circles). (**b**) Frequency plots for the GFN2-xTB+D4 binding energies of both molecular subsets discussed in panel (**a**). We also show the plot corresponding to the energy values obtained using PBE+D3 level of theory for the 1,836 most energetically stable dimer conformation (blue silhouette).

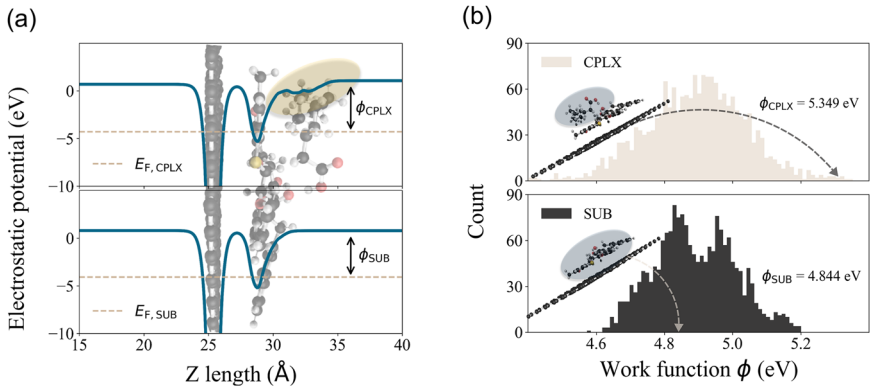

Fig. 5 Understanding work function ϕ calculations of molecular systems in MORE-Q-G3. (**a**) Example case of the system including BOV-35 and Receptor-18: planar-average electrostatic potential along the slab length (z direction) for the CPLX and SUB systems (darkblue curve) with their corresponding Fermi level (dashed line). The arrows for ϕ_{CPLX} and ϕ_{SUB} denote the work function values for each system. The geometry of the CPLX system was placed to correlate the atomic positions with the potential. The change in the potential curve originated from the adsorption of BOV molecule is highlighted by the orange shadow. (**b**) Frequency plots for the work function values of the 1,836 CPLX (white) and SUB (black) systems. The ϕ value for the inserted conformations of CPLX and SUB systems are 5.35 eV and 4.84 eV, respectively.

sample. Additionally, we computed E_{int} at the PBE+D3 level of theory for the 1,836 dimers, and the corresponding values were very similar (see plot with blue solid lines in Fig. 4(b)).

Another unique feature of the MORE-Q dataset is the calculation of the work function, ϕ, for the SUB and CPLX systems (see Eqs. (1)~(4) in "Methods" section). ϕ and the change in work function upon adsorption, $\Delta\phi$, are relevant factors in evaluating the sensing performance *via* the analysis of the electronic responses[80–86]. Indeed, $\Delta\phi$ has already been used as an indicator for the detection of halogen ions in self-assembled alkylammonium halide monolayer-modified substrates[87]. Li *et al.*[88] also conducted systematic density functional theory (DFT) calculations on ten small molecules on the WO_3 substrate, evaluating their sensitivity and selectivity using $\Delta\phi$. Similarly, recent computational studies have highlighted the critical role of work function calculations in sensor applications[89–95], underscoring the significance of this electronic property in molecular sensing. Thus, in our work, ϕ for SUB and CPLX systems was calculated with a careful treatment of the slab thickness and dipole correction, which ensures the smooth and flat region of the electrostatic potential in vacuum E_V, see Fig. 5(a). ϕ is then defined as the energy difference between E_V and Fermi level E_F, see Eq. (1). In the top panel of Fig. 5(a), we show an example of the effect of the adsorption of the BOV molecule on the electrostatic potential of the SUB system (see orange shadow). This alteration of the potential can result in an increase or decrease in ϕ, a magnitude that will be defined as $\Delta\phi$. Figure 5(b) presents the frequency plots of ϕ for SUB and CPLX systems. Here, we observe that the functionalization of the receptor on graphene spreads ϕ in the range [4.6, 5.2] eV, with the highest concentration of ϕ values in the range [4.8, 5.0] eV. With the adsorption of BOV molecules onto the

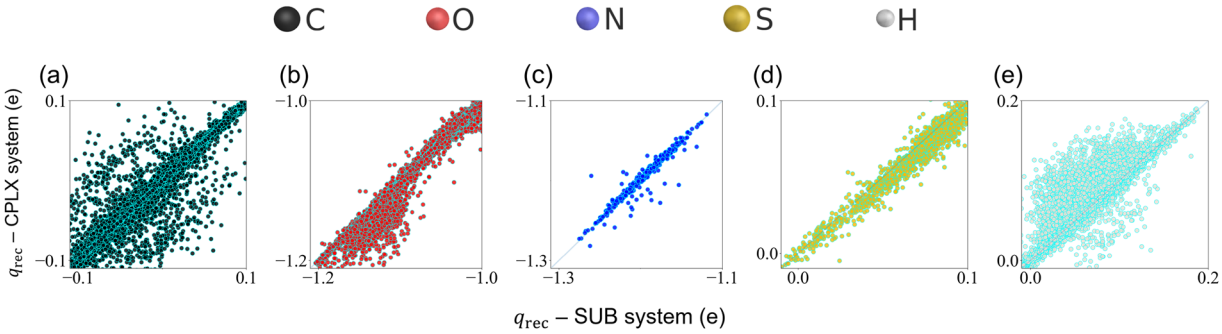

Fig. 6 Correlation plots of the atomic charges in the molecular receptors q_{rec} of CPLX and SUB systems contained in MORE-Q-G3. From (**a**)→(**e**), one can see how atomic charges of Carbon, Oxygen, Nitrogen, Sulfur and Hydrogen atoms in the molecular receptor are modified by the adsorption of BOV molecule.

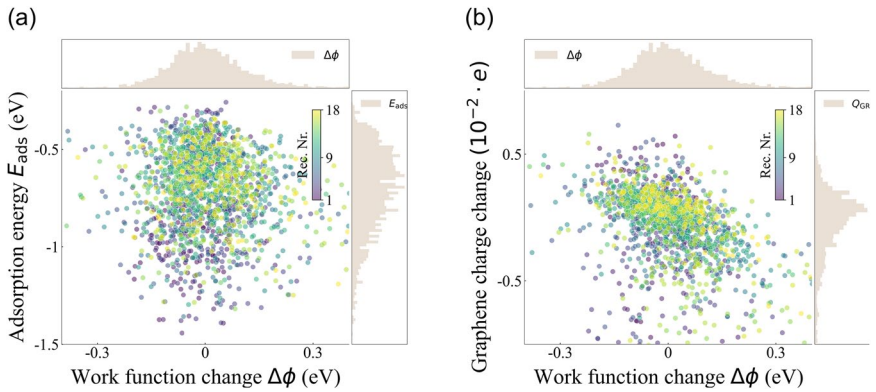

Fig. 7 Validating the change of binding features in systems contained in MORE-Q-G3. Correlation plots between (**a**) adsorption energy E_{ads} and work function change $\Delta\phi$, and (**b**) Graphene charge change and work function change $\Delta\phi$. The corresponding frequency plot per binding feature is also plotted on each panel. For both panels, the datapoints are colored by the receptor number, see the inserted color bar.

SUB systems to form the CPLX system, the bimodal-like distribution transforms into a normal-like distribution, with ϕ values spreading more broadly over the range [4.4, 5.4] eV. This exemplifies the effect of BOV-receptor intermolecular interactions and illustrates the pivotal role of ϕ in understanding the sensing performance of these molecular systems. As shown above, the generation of the MORE-Q dataset involved a series of computational tasks, including GFN-xTB+D4 geometry optimizations, molecular docking conformational searches, and electronic property calculations for both periodic and gas-phase systems. Out of approximately 9 million CPU hours spent performing these calculations, the work function calculations for the 1,836 CPLX and SUB systems were the most computationally expensive, accounting for circa 5 million CPU hours. All calculations were performed on CPUs with Intel Xeon Platinum 8470 processors.

The atomic charges, q, of molecular receptors can also provide insights into sensing performance by examining how the adsorption of BOV molecules (*i.e.* CPLX system) affects the charge distribution in the molecular receptors. As shown in Fig. 6, the atomic charges of the C, H, and O atoms in the receptors are more affected by the BOV-receptor interaction compared to corresponding values of S and N atoms. Moreover, the increase and decrease in (positive/negative) q values reveal the existence of multiple charge transfer processes at reactive atomic sites in the receptors. These changes in charge distribution imply that the local chemical environments in the receptors have been modified — a result that can be further exploited to analyze the sensing mechanisms in these systems.

The change of property values between CPLX and SUB systems is defined as binding features in the present work (see details in "Methods" section). Binding features serve as crucial indicators for the sensing response of SUB systems toward the adsorption of BOV molecules. Figure 7 depicts the correlation plots between select binding features: adsorption energy E_{ads} (see Eq. (5)), graphene charge change ΔQ_{GR}, and work function change $\Delta\phi$. In Fig. 7(a), it can be seen that E_{ads} and $\Delta\phi$ are uncorrelated, varying from -0.3 to -1.5 eV and from -0.3 to 0.3 eV, respectively. Similarly, the change in the atomic charges of the graphene layer does not show any correlation with $\Delta\phi$ (see Figure 7(b)). The small magnitude of the charge changes on graphene (-0.005 to 0.005 e) confirms the weak adsorption of the BOV molecules, while the atomic charge analysis on molecular receptors demonstrates their crucial role in the sensing workforce (*vide supra*). In both panels, the data points are colored according to the receptor number, indicating a lack of clustering based on the receptor. Thus, the

"freedom of design" principle[96] can also be applied to the MORE-Q dataset to gain a better understanding of structure-property and property-property relationships in these potential olfactory sensors.

Since MORE-Q is the first extensive QM dataset for molecular olfactorial receptor engineering, it is expected to have certain limitations. For example, MORE-Q currently spans the chemical space defined by BOV-receptor systems containing only C, H, O, N, and S atoms. To further expand its scope of applicability to more complex BOV molecules related to human body, atom types such as B, F, P, Se, Si, and Ge should be considered[6]. This expansion would also broaden the chemical design space for molecular receptors. Another crucial improvement for MORE-Q involves the inclusion of additional conformations of BOV-receptor systems in the calculations of binding features to construct MORE-Q-G3. This would provide a better understanding of the conformational effects on the olfactory response of molecular receptors. Finally, the computational accuracy of binding feature calculations could be enhanced by using more robust QM methods (*e.g.* hybrid functionals) that include van der Waals corrections, such as D4 or many-body dispersion. However, these improvements would significantly increase computational costs, potentially limiting the number of conformations that can be studied.

In summary, the MORE-Q dataset provides an opportunity to accurately investigate the sensing mechanisms of diverse BOV-receptor systems from a QM perspective. Besides global and local QM properties of monomers and dimer configurations, MORE-Q focuses on relevant binding features such as adsorption energy, charge transfer, and work function changes on graphene substrates. This exhaustive set of QM properties has the potential to enhance the fundamental understanding of the adsorption behavior of BOV molecules. Indeed, by leveraging this knowledge, one can develop robust and transferable machine learning models to predict binding features, which enable a rapid evaluation of sensing performance in molecular systems. MORE-Q can also facilitate the optimization and design of novel mucin-derived molecular receptors through the combination of computed QM property data and generative models, moving closer to the goal of developing biomimetic electronic noses. Additionally, integrating MORE-Q into perception studies of BOV molecules by linking the binding feature space with the human perceptual rating space has the potential to offer valuable insights into the cognitive processes underlying human olfaction.

Code availability

The initial BOV information processing of the screening procedure was conducted mainly using RDkit 2023.09.5[97,98]. Structural optimization has been conducted in the xTB version 6.6.0[59] and DFTB+[67] codes. The configuration search process was performed by the aISS program[60]. The planar-average potential for work function calculation was obtained by Vaspkit[99]. The PBE+D3 property calculation was conducted in ORCA version 5.0.3[68] and VASP version 6.3.1[73,74] packages together with ASE[100]. A user guide for reading and exploring different uses of the MORE-Q dataset has been added to the GitHub repository Repo-MORE-Q[101].

Received: 11 September 2024; Accepted: 11 February 2025;
Published online: 22 February 2025

References

1. Zhang, J., Yin, Z., Chen, P. & Nichele, S. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Inform. fusion* **59**, 103–126, https://doi.org/10.1016/j.inffus.2020.01.011 (2020).
2. Farraia, M. V. *et al*. The electronic nose technology in clinical diagnosis: A systematic review. *Porto Biomed. J.* **4**, e42, https://doi.org/10.1097/j.pbj.0000000000000042 (2019).
3. Mohd Ali, M., Hashim, N., Abd Aziz, S. & Lasekan, O. Principles and recent advances in electronic nose for quality inspection of agricultural and food products. *Trends Food Sci. Technol.* **99**, 1–10, https://doi.org/10.1016/j.tifs.2020.02.028 (2020).
4. Amann, A. *et al*. The human volatilome: volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva. *J. Breath Res.* **8**, 034001, https://doi.org/10.1088/1752-7155/8/3/034001 (2014).
5. Shirasu, M. & Touhara, K. The scent of disease: volatile organic compounds of the human body related to disease and disorder. *J. Biochem.* **150**, 257–266, https://doi.org/10.1093/jb/mvr090 (2011).
6. Drabińska, N. *et al*. A literature survey of all volatiles from healthy human breath and bodily fluids: the human volatilome. *J. Breath Res.* **15**, 034001, https://doi.org/10.1088/1752-7163/abf1d0 (2021).
7. Olsson, M. J. *et al*. The scent of disease: Human body odor contains an early chemosensory cue of sickness. *Psychol. Sci.* **25**, 817–823, https://doi.org/10.1177/0956797613515681 (2014).
8. Buljubasic, F. & Buchbauer, G. The scent of human diseases: a review on specific volatile organic compounds as diagnostic biomarkers. *Flavour Fragrance J.* **30**, 5–25, https://doi.org/10.1002/ffj.3219 (2014).
9. Rondanelli, M. *et al*. Volatile organic compounds as biomarkers of gastrointestinal diseases and nutritional status. *J. Anal. Methods Chem.* **2019**, 1–14, https://doi.org/10.1155/2019/7247802 (2019).
10. Arasaradnam, R. P., Covington, J. A., Harmston, C. & Nwokolo, C. U. Review article: next generation diagnostic modalities in gastroenterology - gas phase volatile compound biomarker detection. *Aliment. Pharmacol. Ther.* **39**, 780–789, https://doi.org/10.1111/apt.12657 (2014).
11. Belizário, J. E., Faintuch, J. & Malpartida, M. G. Breath biopsy and discovery of exclusive volatile organic compounds for diagnosis of infectious diseases. *Front. Cell. Infect. Microbiol.* **10**, https://doi.org/10.3389/fcimb.2020.564194 (2021).
12. Bodelier, A. G. L. *et al*. Volatile organic compounds in exhaled air as novel marker for disease activity in Crohn's disease: A metabolomic approach. *Inflamm. Bowel Dis.* **21**, 1776–1785, https://doi.org/10.1097/mib.0000000000000436 (2015).
13. Sethi, S., Nanda, R. & Chakraborty, T. Clinical application of volatile organic compound analysis for detecting infectious diseases. *Clin. Microbiol. Rev.* **26**, 462–475, https://doi.org/10.1128/cmr.00020-13 (2013).
14. Tisch, U. *et al*. Detection of Alzheimer's and Parkinson's disease from exhaled breath using nanomaterial-based sensors. *Nanomedicine* **8**, 43–56, https://doi.org/10.2217/nnm.12.105 (2012).
15. Bach, J.-P. *et al*. Measuring compounds in exhaled air to detect Alzheimer's disease and Parkinson's disease. *PLoS ONE* **10**, e0132227, https://doi.org/10.1371/journal.pone.0132227 (2015).
16. Trivedi, D. K. *et al*. Discovery of volatile biomarkers of parkinson's disease from sebum. *ACS Cent. Sci.* **5**, 599–606, https://doi.org/10.1021/acscentsci.8b00879 (2019).
17. Havelund, J., Heegaard, N., Færgeman, N. & Gramsbergen, J. Biomarker research in Parkinson's disease using metabolite profiling. *Metabolites* **7**, 42, https://doi.org/10.3390/metabo7030042 (2017).

18. Morgan, J. Joy of super smeller: sebum clues for PD diagnostics. *Lancet Neurol.* **15**, 138–139, https://doi.org/10.1016/s1474-4422(15)00396-8 (2016).
19. Buck, L. & Axel, R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**, 175–187, https://doi.org/10.1016/0092-8674(91)90418-x (1991).
20. Buck, L. B. The molecular architecture of odor and pheromone sensing in mammals. *Cell* **100**, 611–618, https://doi.org/10.1016/s0092-8674(00)80698-4 (2000).
21. Firestein, S. How the olfactory system makes sense of scents. *Nature* **413**, 211–218, https://doi.org/10.1038/35093026 (2001).
22. Garg, N. *et al.* FlavorDB: a database of flavor molecules. *Nucleic Acids Res.* **46**, D1210–D1216, https://doi.org/10.1093/nar/gkx957 (2017).
23. Kumar, Y. *et al.* AromaDb: A database of medicinal and aromatic plant's aroma molecules with phytochemistry and therapeutic potentials. *Front. Plant Sci.* **9**, https://doi.org/10.3389/fpls.2018.01081 (2018).
24. Kumari, S. *et al.* EssOilDB: a database of essential oils reflecting terpene composition and variability in the plant kingdom. *Database* 2014, https://doi.org/10.1093/database/bau120 (2014).
25. Lemfack, M. C. *et al.* mVOC 2.0: a database of microbial volatiles. *Nucleic Acids Res.* **46**, D1261–D1265, https://doi.org/10.1093/nar/gkx1016 (2017).
26. Knudsen, J. T., Eriksson, R., Gershenzon, J. & Ståhl, B. Diversity and distribution of floral scent. *Bot. Rev.* **72**, 1–120, https://doi.org/10.1663/0006-8101(2006)72[1:dadofs]2.0.co;2 (2006).
27. Dunkel, M. *et al.* SuperScent–a database of flavors and scents. *Nucleic Acids Res.* **37**, D291–D294, https://doi.org/10.1093/nar/gkn695 (2009).
28. Gabler, S., Soelter, J., Hussain, T., Sachse, S. & Schmuker, M. Physicochemical vs. vibrational descriptors for prediction of odor receptor responses. *Mol. Inf.* **32**, 855–865, https://doi.org/10.1002/minf.201300037 (2013).
29. Arn, H. & Acree, T. Flavornet: A database of aroma compounds based on odor potency in natural products. *Dev. Food Sci.* **40**, 27–28 (1998).
30. Burns, J. W. & Rogers, D. M. QuantumScents: Quantum-mechanical properties for 3.5k olfactory molecules. *J. Chem. Inf. Model.* **63**, 7330–7337, https://doi.org/10.1021/acs.jcim.3c01338 (2023).
31. Billesbølle, C. B. *et al.* Structural basis of odorant recognition by a human odorant receptor. *Nature* **615**, 742–749, https://doi.org/10.1038/s41586-023-05798-y (2023).
32. Bakhatan, Y. *et al.* Accelerated solid phase glycan synthesis: ASGS. *Chem. - Eur. J.* **29**, https://doi.org/10.1002/chem.202300897 (2023).
33. Olender, T., Nativ, N. & Lancet, D.*HORDE: Comprehensive Resource for Olfactory Receptor Genomics*, 23-38 (Humana Press, 2013).
34. Bateman, A. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489, https://doi.org/10.1093/nar/gkaa1100 (2020).
35. Kwon, O. S., Song, H. S., Park, T. H. & Jang, J. Conducting nanomaterial sensor using natural receptors. *Chem. Rev.* **119**, 36–93, https://doi.org/10.1021/acs.chemrev.8b00159 (2018).
36. Liu, X. *et al.* ODORactor: a web server for deciphering olfactory coding. *Bioinformatics* **27**, 2302–2303, https://doi.org/10.1093/bioinformatics/btr385 (2011).
37. Modena, D., Trentini, M., Corsini, M., Bombaci, A. & Giorgetti, A. OlfactionDB: A database of olfactory receptors and their ligands. *Adv. Life Sci.* **1**, 1–5, https://doi.org/10.5923/j.als.20110101.01 (2012).
38. Mainland, J. D., Li, Y. R., Zhou, T., Liu, W. L. L. & Matsunami, H. Human olfactory receptor responses to odorants. *Sci. Data* **2**, https://doi.org/10.1038/sdata.2015.2 (2015).
39. Sharma, A., Saha, B. K., Kumar, R. & Varadwaj, P. K. Olfactionbase: a repository to explore odors, odorants, olfactory receptors and odorant-receptor interactions. *Nucleic Acids Res.* **50**, D678–D686, https://doi.org/10.1093/nar/gkab763 (2021).
40. Lalis, M. *et al.* M2OR: a database of olfactory receptor-odorant pairs for understanding the molecular mechanisms of olfaction. *Nucleic Acids Res.* **52**, D1370–D1379, https://doi.org/10.1093/nar/gkad886 (2023).
41. Marenco, L. *et al.* ORDB, HORDE, ODORactor and other on-line knowledge resources of olfactory receptor-odorant interactions. *Database* **2016**, baw132, https://doi.org/10.1093/database/baw132 (2016).
42. Audouze, K. *et al.* Identification of odorant-receptor interactions by global mapping of the human odorome. *PLoS ONE* **9**, e93037, https://doi.org/10.1371/journal.pone.0093037 (2014).
43. Triller, A. *et al.* Odorant-receptor interactions and odor percept: A chemical perspective. *Chem. Biodiversity* **5**, 862–886, https://doi.org/10.1002/cbdv.200890101 (2008).
44. Hoja, J. *et al.* QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Sci. Data* **8**, 43, https://doi.org/10.1038/s41597-021-00812-2 (2021).
45. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **4**, 170193, https://doi.org/10.1038/sdata.2017.193 (2017).
46. Montavon, G. *et al.* Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15**, 095003, https://doi.org/10.1088/1367-2630/15/9/095003 (2013).
47. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, https://doi.org/10.1038/sdata.2014.22 (2014).
48. Medrano Sandonas, L. *et al.* Dataset for quantum-mechanical exploration of conformers and solvent effects in large drug-like molecules. *Sci. Data* **11**, 742, https://doi.org/10.1038/s41597-024-03521-8 (2024).
49. Isert, C., Atz, K., Jiménez-Luna, J. & Schneider, G. QMugs, quantum mechanical properties of drug-like molecules. *Sci. Data* **9**, 273, https://doi.org/10.1038/s41597-022-01390-7 (2022).
50. Rezác, J., Riley, K. E. & Hobza, P. Extensions of the S66 data set: More accurate interaction energies and angular-displaced nonequilibrium geometries. *J. Chem. Theory Comput.* **7**, 3466–3470, https://doi.org/10.1021/ct200523a (2011).
51. Sparrow, Z. M., Ernst, B. G., Joo, P. T., Lao, K. U. & DiStasio, R. A. NENCI-2021. I. A large benchmark database of non-equilibrium non-covalent interactions emphasizing close intermolecular contacts. *J. Chem. Phys.* **155**, 184303, https://doi.org/10.1063/5.0068862 (2021).
52. Donchev, A. G. *et al.* Quantum chemical benchmark databases of gold-standard dimer interaction energies. *Sci. Data* **8**, 55, https://doi.org/10.1038/s41597-021-00833-x (2021).
53. Spronk, S. A., Glick, Z. L., Metcalf, D. P., Sherrill, C. D. & Cheney, D. L. A quantum chemical interaction energy dataset for accurately modeling protein-ligand interactions. *Sci. Data* **10**, 619, https://doi.org/10.1038/s41597-023-02443-1 (2023).
54. Geim, A. K. Graphene: status and prospects. *Science* **324**, 1530–1534, https://doi.org/10.1126/science.1158877 (2009).
55. Novoselov, K. S. *et al.* A roadmap for graphene. *Nature* **490**, 192–200, https://doi.org/10.1038/nature11458 (2012).
56. Dai, H. Carbon nanotubes: opportunities and challenges. *Surf. Sci.* **500**, 218–241, https://doi.org/10.1016/s0039-6028(01)01558-8 (2002).
57. Popov, V. Carbon nanotubes: properties and application. *Mater. Sci. Eng. R Rep.* **43**, 61–102, https://doi.org/10.1016/j.mser.2003.10.001 (2004).
58. Tasis, D., Tagmatarchis, N., Bianco, A. & Prato, M. Chemistry of carbon nanotubes. *Chem. Rev.* **106**, 1105–1136, https://doi.org/10.1021/cr050569o (2006).

59. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB-an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **15**, 1652–1671, https://doi.org/10.1021/acs.jctc.8b01176 (2019).
60. Plett, C. & Grimme, S. Automated and efficient generation of general molecular aggregate structures. *Angew. Chem., Int. Ed.* **62**, e202214477, https://doi.org/10.1002/anie.202214477 (2023).
61. Ben Abba Amiel, D. & Hurevich, M. Expeditious synthesis of a glycopeptide library. *Eur. J. Org. Chem.* **2022**, e202200623, https://doi.org/10.1002/ejoc.202200623 (2022).
62. Sukhran, Y. *et al*. Unexpected nucleophile masking in acyl transfer to sterically crowded and conformationally restricted galactosides. *J. Org. Chem.* **88**, 9313–9320, https://doi.org/10.1021/acs.joc.3c00878 (2023).
63. Caldeweyher, E. *et al*. A generally applicable atomic-charge dependent London dispersion correction. *J. Chem. Phys.* 150, https://doi.org/10.1063/1.5090222 (2019).
64. Keller, A. & Vosshall, L. B. Olfactory perception of chemically diverse molecules. *BMC Neurosci.* **17**, 1–17, https://doi.org/10.1186/s12868-016-0287-2 (2016).
65. Kim, S. *et al*. Pubchem 2023 update. *Nucleic acids* **51**, D1373–D1380, https://doi.org/10.1093/nar/gkac956 (2023).
66. Hourahine, B. *et al*. DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *J. Chem. Phys.* 152, https://doi.org/10.1063/1.5143190 (2020).
67. Aradi, B., Hourahine, B. & Frauenheim, T. DFTB+, a sparse matrix-based implementation of the DFTB method. *J. Phys. Chem. A* **111**, 5678–5684, https://doi.org/10.1021/jp070186p (2007).
68. Neese, F. The ORCA program system. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2**, 73–78, https://doi.org/10.1002/wcms.81 (2012).
69. Mulliken, R. S. Electronic population analysis on LCAO–MO molecular wave functions. I. *J. Chem. Phys* **23**, 1833–1840, https://doi.org/10.1063/1.1740588 (1955).
70. Löwdin, P.-O. On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals. *J. Chem. Phys* **18**, 365–375, https://doi.org/10.1063/1.1747632 (1950).
71. Mayer, I. Charge, bond order and valence in the AB initio SCF theory. *Chem. Phys. Lett.* **97**, 270–274, https://doi.org/10.1016/0009-2614(83)80005-0 (1983).
72. Grimme, S., Bannwarth, C., Caldeweyher, E., Pisarek, J. & Hansen, A. A general intermolecular force field based on tight-binding quantum chemical calculations. *J. Chem. Phys.***147**, https://doi.org/10.1063/1.4991798 (2017).
73. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169, https://doi.org/10.1103/PhysRevB.54.11169 (1996).
74. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50, https://doi.org/10.1016/0927-0256(96)00008-0 (1996).
75. Henkelman, G., Arnaldsson, A. & Jónsson, H. A fast and robust algorithm for Bader decomposition of charge density. *Comput. Mater. Sci.* **36**, 354–360, https://doi.org/10.1016/j.commatsci.2005.04.010 (2006).
76. Chen, L. *et al*. Datasets and geometries for "MORE-Q, Dataset for molecular olfactorial receptor engineering by quantum mechanics", *ZENODO*, https://doi.org/10.5281/ZENODO.13741197 (2024).
77. Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465, https://doi.org/10.1002/jcc.21759 (2011).
78. Plett, C., Grimme, S. & Hansen, A. Conformational energies of biomolecules in solution: Extending the MPCONF196 benchmark with explicit water molecules. *J. Comput. Chem.* **45**, 419–429, https://doi.org/10.1002/jcc.27248 (2024).
79. Murtagh, F. & Contreras, P. Algorithms for hierarchical clustering: an overview. *WIREs Data. Mining. Knowl. Discov.* **2**, 86–97, https://doi.org/10.1002/widm.53 (2011).
80. Lin, L. *et al*. Work function: Fundamentals, measurement, calculation, engineering, and applications. *Phys. Rev. Appl.* **19**, https://doi.org/10.1103/physrevapplied.19.037001 (2023).
81. Gurlo, A., Sahm, M., Oprea, A., Barsan, N. & Weimar, U. A p- to n-transition on α-Fe_2O_3-based thick film sensors studied by conductance and work function change measurements. *Sens. Actuators, B* **102**, 291–298, https://doi.org/10.1016/j.snb.2004.04.075 (2004).
82. Chou, P.-C. *et al*. On a Schottky diode-type hydrogen sensor with pyramid-like Pd nanostructures. *Int. J. Hydrogen Energy* **40**, 9006–9012, https://doi.org/10.1016/j.ijhydene.2015.05.036 (2015).
83. Kumar, A. Palladium-based trench gate MOSFET for highly sensitive hydrogen gas sensor. *Mater. Sci. Semicond. Process* **120**, 105274, https://doi.org/10.1016/j.mssp.2020.105274 (2020).
84. Pour, G. B., Aval, L. F. & Eslami, S. Sensitive capacitive-type hydrogen sensor based on Ni thin film in different hydrogen concentrations. *Curr. Nanosci.* **14**, 136–142 (2018).
85. Sahm, T., Gurlo, A., Bârsan, N. & Weimar, U. Basics of oxygen and SnO_2 interaction; work function change and conductivity measurements. *Sens. Actuators, B* **118**, 78–83, https://doi.org/10.1016/j.snb.2006.04.004 (2006).
86. Meng, J. & Li, Z. Schottky-contacted nanowire sensors. *Adv. Mater.* **32**, 2000130, https://doi.org/10.1002/adma.202000130 (2020).
87. Gankin, A. *et al*. Molecular and ionic dipole effects on the electronic properties of Si-/SiO_2-grafted alkylamine monolayers. *ACS Appl. Mater. Interfaces* **9**, 44873–44879, https://doi.org/10.1021/acsami.7b12218 (2017).
88. Li, J.-H., Wu, J. & Yu, Y.-X. DFT exploration of sensor performances of two-dimensional WO_3 to ten small gases in terms of work function and band gap changes and I-V responses. *Appl. Surf. Sci.* **546**, 149104, https://doi.org/10.1016/j.apsusc.2021.149104 (2021).
89. Nath, U. & Sarma, M. Pyridinic dominance N-doped graphene: A potential material for SO_2 gas detection. *J. Phys. Chem. A* **127**, 1112–1123, https://doi.org/10.1021/acs.jpca.2c06154 (2023).
90. Cid, B. J. *et al*. Metal-decorated siligene as work function type sensor for NH3 detection: A DFT approach. *Appl. Surf. Sci.* **610**, 155541, https://doi.org/10.1016/j.apsusc.2022.155541 (2023).
91. Kalwar, B. A. *et al*. Highly sensitive work function type room temperature gas sensor based on Ti doped hBN monolayer for sensing CO_2, CO, H_2S, HF and NO. A DFT study. *RSC Advances* **12**, 34185–34199, https://doi.org/10.1039/d2ra06307g (2022).
92. Cui, H., Jia, P., Peng, X. & Li, P. Adsorption and sensing of CO and C_2H_2 by S-defected SnS_2 monolayer for DGA in transformer oil: A DFT study. *Mater. Chem. Phys.* **249**, 123006, https://doi.org/10.1016/j.matchemphys.2020.123006 (2020).
93. Ni, J., Wang, W., Quintana, M., Jia, F. & Song, S. Adsorption of small gas molecules on strained monolayer WSe_2 doped with Pd, Ag, Au, and Pt: A computational investigation. *Appl. Surf. Sci.* **514**, 145911, https://doi.org/10.1016/j.apsusc.2020.145911 (2020).
94. Reji, R. P., Balaji, S. K. C., Sivalingam, Y., Kawazoe, Y. & Velappa Jayaraman, S. First-principles density functional theory calculations on the potential of Sc_2CO_2 MXene nanosheets as a dual-mode sensor for detection of volatile organic compounds in exhaled human breath. *ACS Appl. Nano Mater.* **6**, 5345–5356, https://doi.org/10.1021/acsanm.2c05474 (2023).
95. Lin, L. *et al*. DFT study on the adsorption of CO, NO_2, SO_2 and NH_3 by Te vacancy and metal atom doped $MoTe_2$ monolayers. *Phys. E* **145**, 115489, https://doi.org/10.1016/j.physe.2022.115489 (2023).
96. Medrano Sandonas, L. *et al*. "freedom of design" in chemical compound space: towards rational *in silico* design of molecules with targeted quantum-mechanical properties. *Chem. Sci.* **14**, 10702–10717, https://doi.org/10.1039/d3sc03598k (2023).
97. Landrum, G. *et al*. RDKit: Open-source cheminformatics. Release 2014.03.1, https://doi.org/10.5281/ZENODO.10398 (2014).
98. Landrum, G. *et al*. rdkit/rdkit: Release 2023.09.5, https://doi.org/10.5281/ZENODO.10633624 (2024).
99. Wang, V., Xu, N., Liu, J.-C., Tang, G. & Geng, W.-T. VASPKIT: A user-friendly interface facilitating high-throughput computing and analysis using VASP code. *Comput. Phys. Commun.* **267**, 108033, https://doi.org/10.1016/j.cpc.2021.108033 (2021).

100. Hjorth Larsen, A. *et al*. The atomic simulation environment-a Python library for working with atoms. *J. Condens. Matter Phys.* **29**, 273002, https://doi.org/10.1088/1361-648x/aa680e (2017).
101. Chen, L. *et al*. MORE-Q. GitHub repository https://github.com/LiC1117/MORE-Q (2025).

Acknowledgements

The authors gratefully acknowledge the funding by the European Union Horizon Europe EIC Pathfinder Open project "Smart Electronic Olfaction for Body Odor Diagnostics" (SMELLODI, grant agreement ID: 101046369), the Volkswagen Foundation for the Qualification Concept "Olfactorial Perceptronics" (Project ID 9B396), and Federal Ministry of Education and Research of Germany in the program of "Souverän. Digital. Vernetzt." joint project 6G-life, project number: 16KISK001K. We acknowledge also the Center for Information Services and High Performance Computing (ZIH) at TU Dresden for providing the computational resources.

Author contributions

P.T. selected and analyzed the initial set of BOV molecules. M.H. and S.Y. proposed the newly-synthesised bio-inspired (mucin-derived) receptors. L.C. performed the QM calculations for all structures using the HPC facilities at TU Dresden. L.C. and L.M.S. designed and wrote the manuscript. L.M.S., A.D. and N.T. contributed to the curation and technical validation of the dataset. L.M.S., A.D., R.G., A.C., and G.C. supervised and revised all stages of the work. All authors discussed the results and contributed to the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interest

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04616-6.

Correspondence and requests for materials should be addressed to L.M.S., A.C. or G.C.