Supporting Information for Unifying Digital Discovery of Electronic Materials with EMOS

Atish Dixit[1*] Alexandros Keros[1] Ben D. Rowlinson[1] Shashank Mishra[1]
Jacqueline M. Cole[2] Subramanian Ramamoorthy[3] Themis Prodromakis[1]

S1 EMOS Materials Database Composition

The current release of EMOS aggregates six complementary database *IUs* to provide broad entry-point coverage for electronic materials discovery. Rather than relying on a single repository, EMOS combines large general-purpose sources with smaller but domain-useful collections, which improves practical retrieval breadth across diverse workflow contexts. This multi-source composition is operationally important because it reduces dependence on any one schema, curation policy, or access endpoint, and instead exposes a unified access layer across heterogeneous data providers. As summarised in Figure S1, the integrated set currently provides a total of 1,757,563 accessible entries in total; this value reflects cumulative accessible records across connected sources and should therefore be interpreted as aggregate availability rather than strict de-duplicated unique-material count. The same integration strategy also supports mixed access backends (standardised APIs and source-specific interfaces), which helps maintain extensibility as additional databases are onboarded.

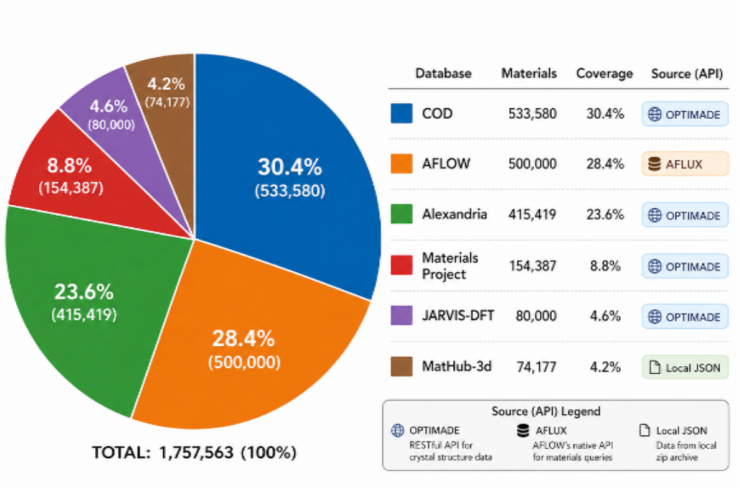


Figure S1: **Database composition snapshot for the current EMOS release.** The pie chart reports each source contribution as both percentage and absolute record count, while the adjacent table lists per-database totals and the corresponding access route (OPTIMADE, AFLUX, or local JSON), including the API legend used in the framework view.

*Corresponding author: adixit@ed.ac.uk

S2 Implementation Details of Database Extractor *Feature*

The Database Extractor is a dedicated EMOS *feature* implemented using the EMOS template generator, which provides the standard scaffolding for all *features* in the framework. The *feature* is illustrated in Figure S2.

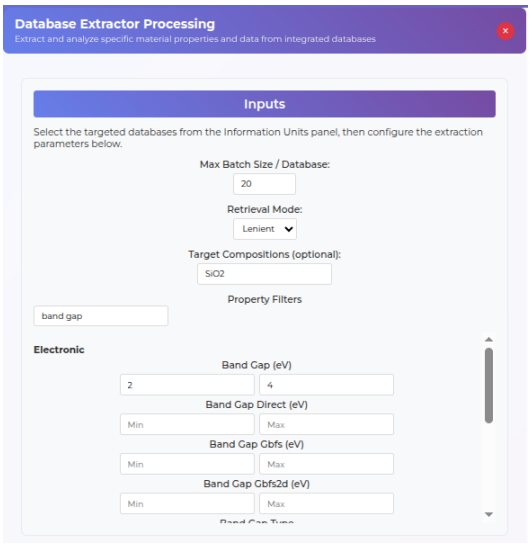
Inputs (Figure S2a). The input panel exposes the following configuration parameters:

- **Information Units (databases):** The user selects one or more databases from the set of integrated EMOS database *IUs*. Only the databases selected here will be queried during extraction.
- **Batch size:** The maximum number of records to retrieve per database per query. This allows users to control the volume of results returned from each source independently.
- **Retrieval mode:** Selects between *strict* mode, in which a material is only returned if all chosen property filters are satisfied across every selected database, and *lenient* mode, in which each database is queried independently based on whichever of the chosen properties it supports, yielding a broader candidate pool.
- **Target composition:** An optional composition constraint (e.g. AlN, SiC) that restricts retrieval to materials with a composition similar to the specified reference.
- **Property filters:** One or more property constraints used to filter the retrieved candidates. The available properties are drawn from the full EMOS property dictionary dynamically mapped to all the available databases. Because this dictionary is large, properties are grouped into categories such as electronic, mechanical, thermodynamic, and structural to aid navigation. A search filter is also provided to help users quickly locate a specific property by name without browsing all categories. Each selected property can be given a numerical range (e.g. band gap: 2.0–6.5 eV), and only materials satisfying the specified range are returned.

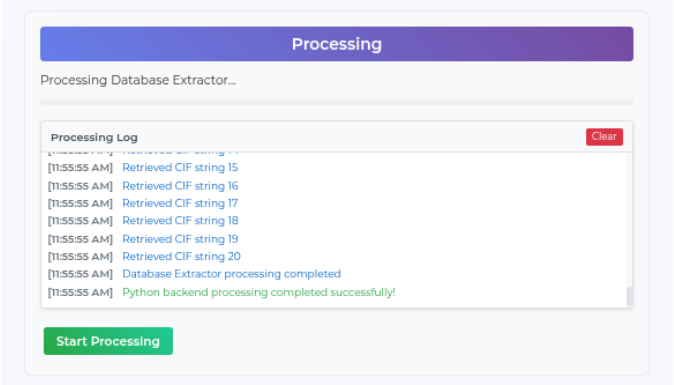
Processing (Figure S2b). Once the inputs have been submitted, the *feature* dispatches parallel queries to all selected databases. A live processing log streams retrieval events in real time, with timestamped entries confirming each CIF record fetched from each queried database as the extraction proceeds. A final confirmation message is displayed upon successful completion of the Python backend.

Outputs (Figure S2c). Upon completion, the *feature* presents a structured extraction summary with the following fields:

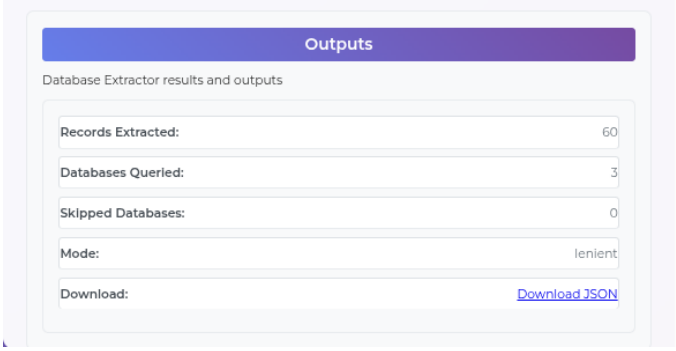
- **Total records extracted:** The total number of candidate materials returned across all selected databases after applying the specified filters.
- **Databases queried:** The number of databases that were successfully queried and returned results.
- **Databases skipped:** Any databases that were skipped during extraction, typically because they do not support one or more of the requested property filters.
- **Retrieval mode used:** The mode (*strict* or *lenient*) applied during the extraction run, confirming the filtering strategy used to assemble the candidate pool.
- **Download:** The full candidate set, including all retrieved CIF records and associated property values, can be downloaded as a JSON (JavaScript Object Notation) file for direct use in downstream workflows or external analysis pipelines.



(a) Input configuration panel, showing an example SiO_2-based query with a band gap filter of 2-4 eV.



(b) Live processing log showing timestamped retrieval events and backend completion confirmation.



(c) Output summary showing 60 records extracted from 3 databases in *lenient* mode, with JSON download.

Figure S2: **Database Extractor user interface in EMOS.** (a) input configuration, (b) processing log, and (c) output summary.

S3 CIF Similarity Supporting Information

S3.1 Tables for list of materials inthe case study

Table S1: Complete ZnO structure dataset retrieved from MP, COD and OQMD using EMOS

file_name	**formula**	**spg**	**spg#**	**sites**	***a***	***b***	***c***	**α**	**β**	**γ**	**system**
cod-1011258	Zn_2 O_2	$P6_3mc$	186	4	3.22	3.22	5.2	90	90	120	hexagonal
cod-1011259	Zn_2 O_2	$P6_3mc$	186	4	3.351	3.351	5.226	90	90	120	hexagonal
cod-1534836	Zn_4 O_4	$Fm\bar{3}m$	225	8	4.28	4.28	4.28	90	90	90	cubic
cod-1537875	Zn_4 O_4	$F\bar{4}3m$	216	8	4.629	4.629	4.629	90	90	90	cubic
cod-2107059	Zn_2 O_2	$P6_3mc$	186	4	3.2417	3.2417	5.1876	90	90	120	hexagonal
cod-2300112	Zn_2 O_2	$P6_3mc$	186	4	3.2494	3.2494	5.2054	90	90	120	hexagonal
cod-2300113	Zn_2 O_2	$P6_3mc$	186	4	3.2342	3.2342	5.1772	90	90	120	hexagonal
cod-2300114	Zn_2 O_2	$P6_3mc$	186	4	3.219	3.219	5.1489	90	90	120	hexagonal
cod-2300115	Zn_2 O_2	$P6_3mc$	186	4	3.2049	3.2049	5.1216	90	90	120	hexagonal
cod-2300116	Zn_2 O_2	$P6_3mc$	186	4	3.195	3.195	5.1027	90	90	120	hexagonal
cod-2300450	Zn_2 O_2	$P6_3mc$	186	4	3.2493	3.2493	5.2057	90	90	120	hexagonal
cod-9004178	Zn_2 O_2	$P6_3mc$	186	4	3.2494	3.2494	5.2038	90	90	120	hexagonal
cod-9004179	Zn_2 O_2	$P6_3mc$	186	4	3.2533	3.2533	5.2073	90	90	120	hexagonal
cod-9004180	Zn_2 O_2	$P6_3mc$	186	4	3.2494	3.2494	5.2038	90	90	120	hexagonal
cod-9004181	Zn_2 O_2	$P6_3mc$	186	4	3.2533	3.2533	5.2073	90	90	120	hexagonal
cod-9008877	Zn_2 O_2	$P6_3mc$	186	4	3.2495	3.2495	5.2069	90	90	120	hexagonal
cod-9011662	Zn_2 O_2	$P6_3mc$	186	4	3.249	3.249	5.207	90	90	120	hexagonal
mp-1017539	Zn_2 O_2	$P6_3mc$	186	4	3.2052	3.2052	5.5169	90	90	120	hexagonal
mp-1093993	Zn_4 O_4	$P4_2/mnm$	136	8	5.5176	5.5176	3.2588	90	90	90	tetragonal
mp-1245015	Zn_{40} O_{40}	P1	1	80	10.4428	9.9768	10.7918	88.8271	84.3724	91.3308	triclinic
mp-1245109	Zn_{40} O_{40}	P1	1	80	9.9377	10.1362	10.5709	90.2642	93.0142	84.3384	triclinic
mp-1245209	Zn_{40} O_{40}	P1	1	80	10.0898	10.7816	9.9926	88.5069	86.8414	89.6777	triclinic
mp-1245255	Zn_{40} O_{40}	P1	1	80	10.6406	10.9388	9.4722	92.6689	87.733	80.6212	triclinic
mp-13161	Zn O	$Pm\bar{3}m$	221	2	2.6919	2.6919	2.6919	90	90	90	cubic
mp-1986	Zn O	$F\bar{4}3m$	216	2	3.2234	3.2234	3.2234	60	60	60	cubic
mp-2133	Zn_2 O_2	$P6_3mc$	186	4	3.2374	3.2374	5.2221	90	90	120	hexagonal
mp-2229	Zn O	$Fm\bar{3}m$	225	2	3.0681	3.0681	3.0681	60	60	60	cubic
mp-2346983	Zn_2 O_2	$P6_3/mmc$	194	4	3.288	3.288	7.1304	90	90	120	hexagonal
mp-997630	Zn_{12} O_{12}	$P6_3/mmc$	194	24	3.0555	3.0555	30.4099	90	90	120	hexagonal
oqmd-1106917	Zn_4 O_4	$Fm\bar{3}m$	225	8	4.3009	4.3009	4.3009	90	90	90	cubic

file_name	formula	spg	spg#	sites	*a*	*b*	*c*	α	β	γ	system
oqmd-116025	Zn_2 O_2	$P6_3/mmc$	194	4	3.0043	3.0043	5.1412	90	90	120	hexagonal
oqmd-117131	Zn O	$Pm\bar{3}m$	221	2	2.6652	2.6652	2.6652	90	90	90	cubic
oqmd-1223683	Zn_4 O_4	$F\bar{4}3m$	216	8	4.5887	4.5887	4.5887	90	90	90	cubic
oqmd-1230655	Zn O	P-6m2	187	2	2.9353	2.9353	2.6942	90	90	120	hexagonal
oqmd-1343236	Zn_4 O_4	$Fm\bar{3}m$	225	8	4.2988	4.2988	4.2988	90	90	90	cubic
oqmd-1343668	Zn_2 O_2	$P4_2/mmc$	131	4	3.002	3.002	5.3962	90	90	90	tetragonal
oqmd-1343724	Zn_2 O_2	I4/mmm	139	4	3.1088	3.1088	4.0981	90	90	90	tetragonal
oqmd-1343728	Zn_2 O_2	P4/nmm	129	4	3.9902	3.9902	3.4567	90	90	90	tetragonal
oqmd-1439997	Zn_4 O_4	$Fm\bar{3}m$	225	8	4.296	4.296	4.296	90	90	90	cubic
oqmd-1442030	Zn_4 O_4	$Fm\bar{3}m$	225	8	4.3008	4.3008	4.3008	90	90	90	cubic
oqmd-1442249	Zn_4 O_4	$P4_2/mnm$	136	8	5.494	5.494	3.3075	90	90	90	tetragonal
oqmd-1443821	Zn_2 O_2	$P6_3/mmc$	194	4	3.4655	3.4655	4.3714	90	90	120	hexagonal
oqmd-1443987	Zn_6 O_6	$Pm\bar{3}n$	223	12	5.6071	5.6071	5.6071	90	90	90	cubic
oqmd-1444341	Zn_4 O_4	$Fm\bar{3}m$	225	8	4.2895	4.2895	4.2895	90	90	90	cubic
oqmd-1472687	Zn_4 O_4	$Fm\bar{3}m$	225	8	4.2996	4.2996	4.2996	90	90	90	cubic
oqmd-1472817	Zn_9 O_9	R3m	160	18	8.8641	8.8641	2.8159	90	90	120	trigonal
oqmd-1473257	Zn_4 O_4	$Fm\bar{3}m$	225	8	4.298	4.298	4.298	90	90	90	cubic
oqmd-1474198	Zn_{12} O_{12}	Cmcm	63	24	3.27	5.774	16.8032	90	90	90	orthorhombic
oqmd-1479504	Zn_8 O_8	$P4_2/m$	84	16	5.6368	5.6368	5.67	90	90	90	tetragonal
oqmd-1479511	Zn_8 O_8	$P4_2/m$	84	16	5.6368	5.6368	5.6701	90	90	90	tetragonal
oqmd-2029561	Zn_2 O_2	I4mm	107	4	2.8933	2.8933	5.0525	90	90	90	tetragonal
oqmd-2029562	Zn_2 O_2	$P6_3/mmc$	194	4	3.4486	3.4486	4.5045	90	90	120	hexagonal
oqmd-2029563	Zn_2 O_2	$P6_3/mmc$	194	4	2.9997	2.9997	5.1608	90	90	120	hexagonal
oqmd-2029564	Zn_4 O_4	$P4_2/mnm$	136	8	5.5657	5.5657	3.2623	90	90	90	tetragonal
oqmd-2029839	Zn_4 O_4	$P6_3mc$	186	8	3.2524	3.2524	10.5691	90	90	120	hexagonal
oqmd-2029840	Zn_5 O_5	P3m1	156	10	3.2509	3.2509	13.2085	90	90	120	trigonal
oqmd-2029841	Zn_6 O_6	$P6_3mc$	186	12	3.248	3.248	15.8737	90	90	120	hexagonal
oqmd-2029842	Zn_8 O_8	$P6_3mc$	186	16	3.2441	3.2441	21.1445	90	90	120	hexagonal
oqmd-2029843	Zn_9 O_9	R3m	160	18	3.2548	3.2548	23.6397	90	90	120	trigonal
oqmd-2029844	Zn_{12} O_{12}	R3m	160	24	3.2498	3.2498	31.7135	90	90	120	trigonal
oqmd-20684	Zn O	$Pm\bar{3}m$	221	2	2.6604	2.6604	2.6604	90	90	90	cubic
oqmd-20685	Zn_4 O_4	$F\bar{4}3m$	216	8	4.5774	4.5774	4.5774	90	90	90	cubic

file_name	**formula**	**spg**	**spg#**	**sites**	***a***	***b***	***c***	**α**	**β**	**γ**	**system**
oqmd-2170474	Zn_6 O_6	P3m1	156	12	3.256	3.256	15.8327	90	90	120	trigonal
oqmd-307732	Zn O	$Pm\bar{3}m$	221	2	2.6662	2.6662	2.6662	90	90	90	cubic
oqmd-328816	Zn_4 O_4	$Fm\bar{3}m$	225	8	4.2903	4.2903	4.2903	90	90	90	cubic
oqmd-39274	Zn O	$Pm\bar{3}m$	221	2	2.6631	2.6631	2.6631	90	90	90	cubic
oqmd-4908	Zn_2 O_2	$P6_3mc$	186	4	3.2599	3.2599	5.2214	90	90	120	hexagonal
oqmd-6958	Zn_4 O_4	$Fm\bar{3}m$	225	8	4.2918	4.2918	4.2918	90	90	90	cubic

S3.2 Effect of the choice of neighbourhood index k on AMD analysis

S3.2.1 Local Neighbourhood (k=5)

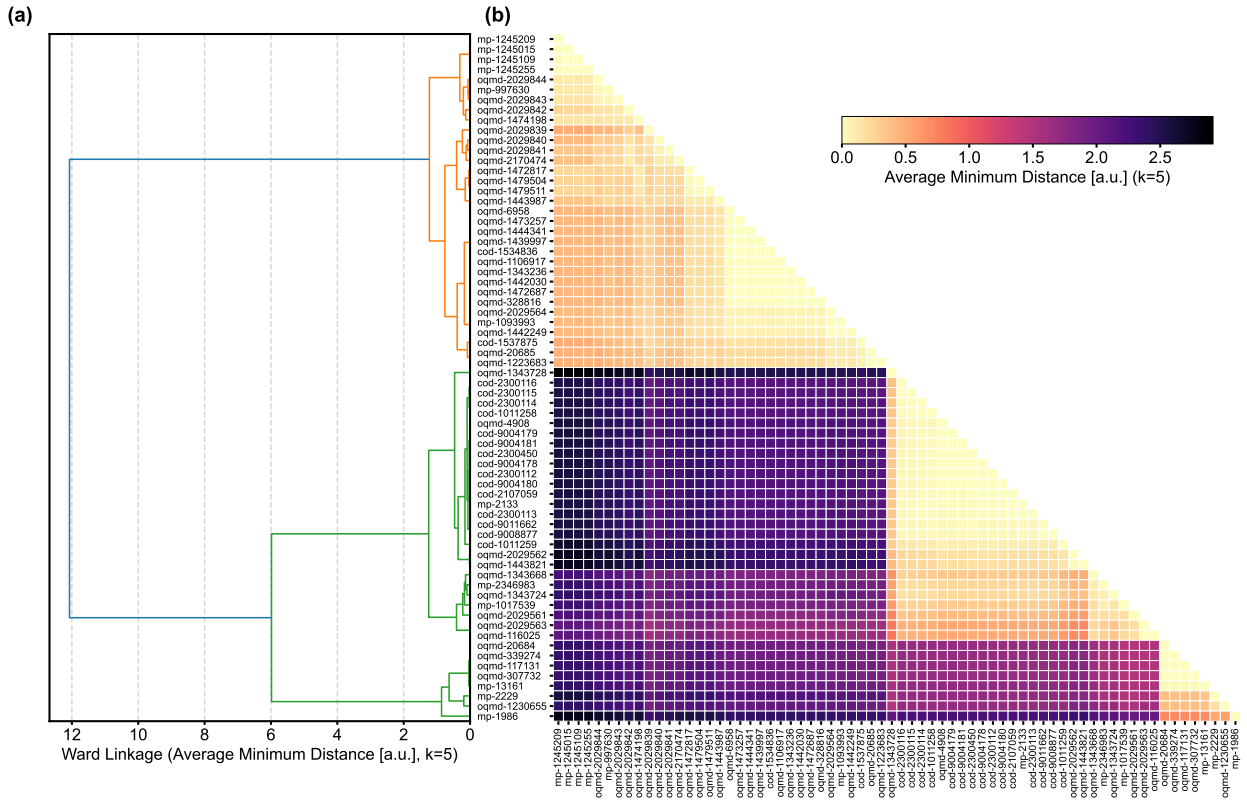


Figure S3: Neighbourhood Index $k = 5$

S3.2.2 k=25

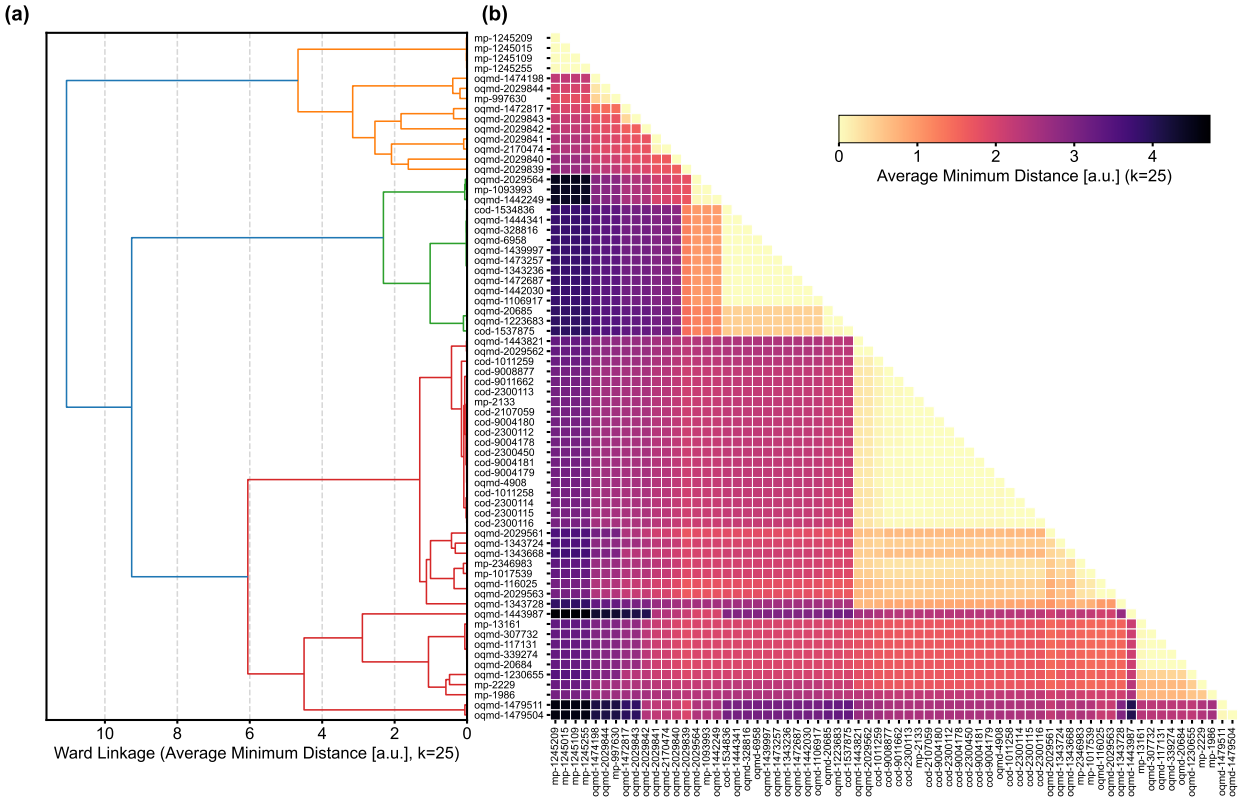


Figure S4: Neighbourhood Index $k = 25$

S3.2.3 k=50

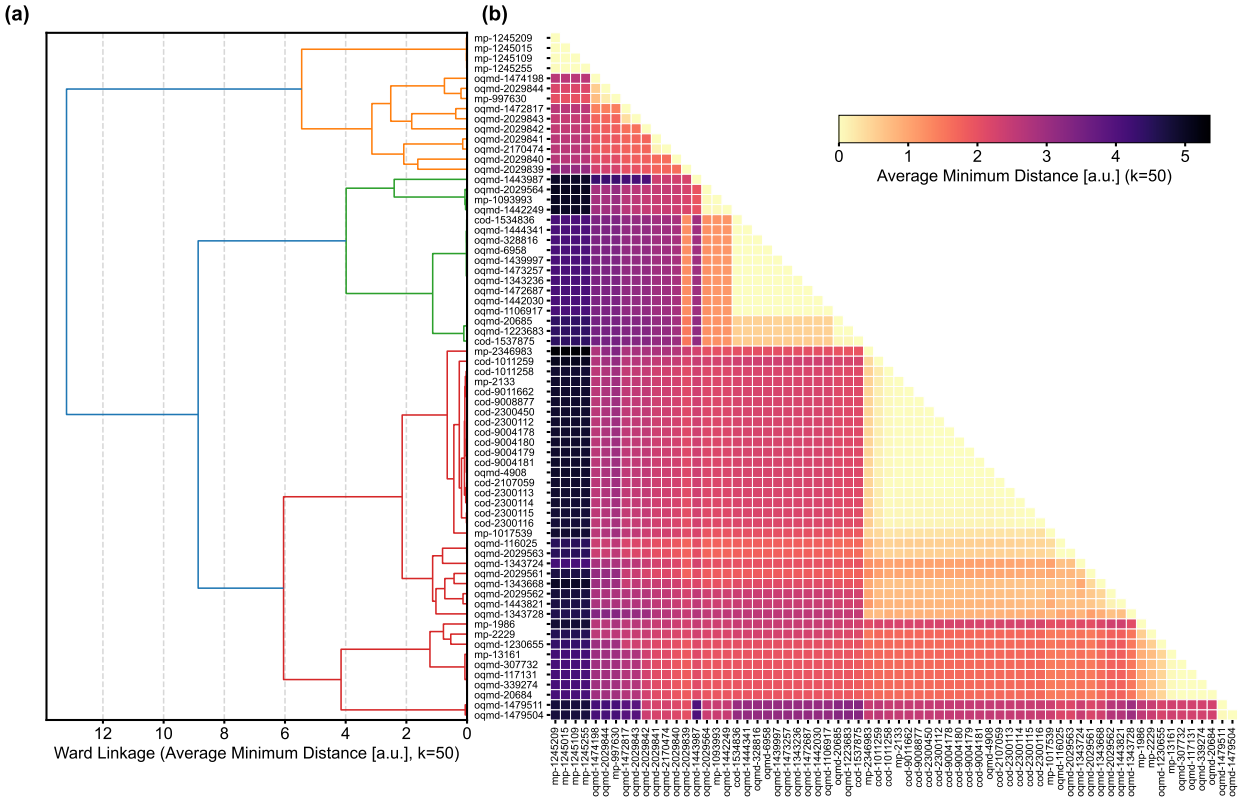


Figure S5: Neighbourhood Index $k = 50$

S3.2.4 Supercell (k=100)

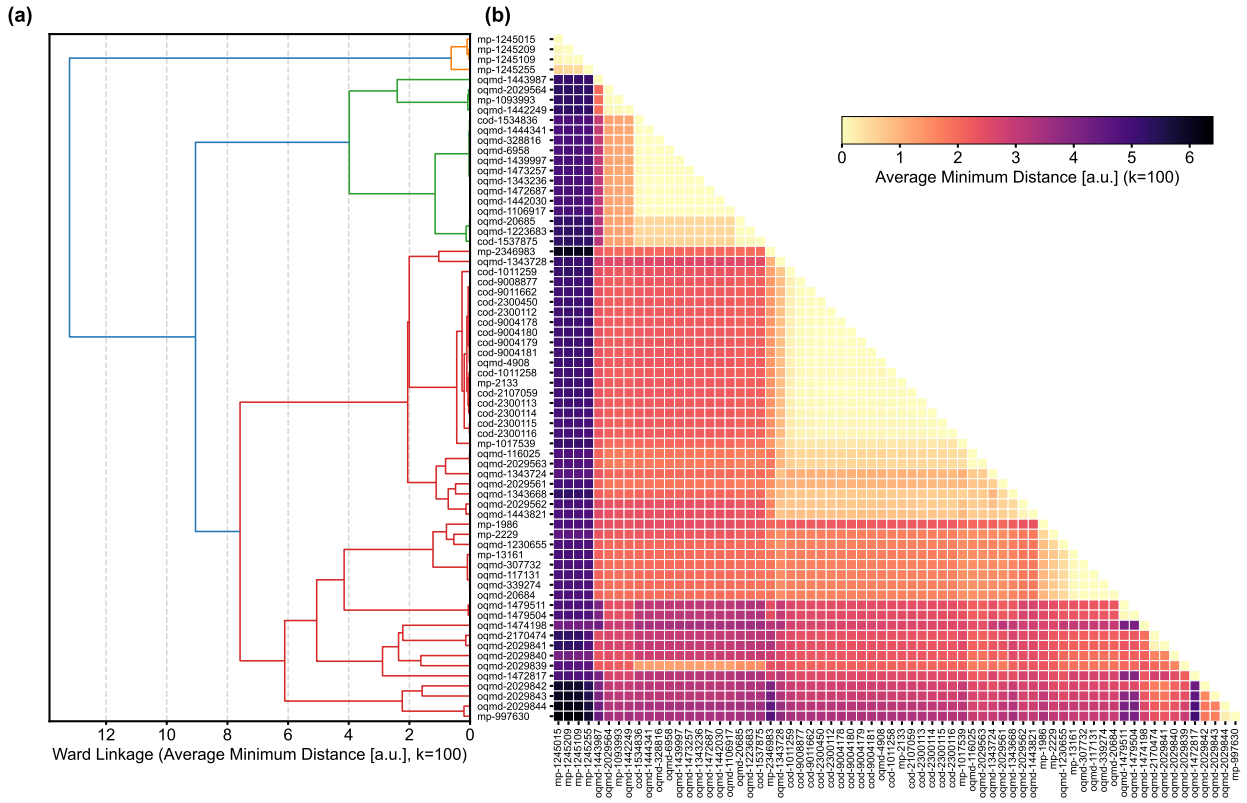


Figure S6: Neighbourhood Index $k = 100$

S4 MOSFET Material Evaluation: Model Description and Default Parameters

Model Overview

Device-level evaluation is performed with a two-dimensional (2D) Poisson–drift-diffusion (PDD) solver implemented in Python, ported from the educational MATLAB drift-diffusion code by Chien-Ting Tung (UC Berkeley) [45]. For each gate–drain bias point $(V_{\mathrm{gs}}, V_{\mathrm{ds}})$, the solver self-consistently advances the Poisson equation,

$$\nabla \cdot \big(\varepsilon \, \nabla \phi\big) = -q\big(p - n + N_B\big), \tag{S1}$$

and the electron and hole carrier-continuity equations until the maximum change in the conduction-band edge between successive iterations falls below 10^{-6}. Here, ε denotes the material permittivity, ϕ is the electrostatic potential, q is the elementary charge, n and p are the electron and hole carrier concentrations, respectively, and N_B represents the net background doping concentration.

Carrier transport employs a field-dependent mobility model with velocity saturation,

$$\mu_{\mathrm{eff}} = \frac{\mu_0}{\left(1 + \left(\dfrac{|\nabla E_c| \, \mu_0}{q \, v_{\mathrm{sat}}}\right)^{p}\right)^{1/p}}, \tag{S2}$$

where μ_{eff} is the effective field-dependent mobility, μ_0 is the low-field mobility, E_c is the conduction-band edge energy (with $|\nabla E_c|$ representing the driving electric field magnitude), v_{sat} is the carrier saturation velocity, and p is the material-specific velocity-saturation power-law exponent. The drain current per unit width is obtained by integrating the electron and hole current densities at the drain contact over the channel cross-section.

Device Geometry and Bias Conditions

The device is a double-gate n-channel MOSFET in which the channel layer is sandwiched symmetrically between two insulator layers of equal thickness, with gate bias applied identically at both interfaces. Table S2 lists the default geometry, contact, and bias parameters used throughout the screening workflow.

Default Gate Insulator Parameters

The SiO_2-reference defaults shown in Table S3 are held fixed across all candidate evaluations.

Device Screening and Ranking Criteria

Candidates are accepted only if they satisfy both of the following thresholds:

$$I_{\mathrm{off}} \le 10^{-2} \ \mu\mathrm{A}/\mu\mathrm{m}, \qquad 0.15 \ \mathrm{V} \le V_{\mathrm{th}} \le 0.65 \ \mathrm{V}. \tag{S3}$$

Accepted candidates are ranked in the following priority order:

1. $I_{\mathrm{on}}/I_{\mathrm{off}}$ ratio — descending (primary metric);
2. I_{on} — descending;
3. $|V_{\mathrm{th}} - 0.35 \ \mathrm{V}|$ — ascending.

Table S2: Default device geometry and bias parameters.

Parameter	Default value	Unit
Channel length	14	nm
Source/drain length	4	nm
Gate oxide thickness (each side)	1	nm
Channel thickness	4	nm
Mesh spacing (x and y)	0.5	nm
Temperature	300	K
Gate work function	3.65	eV
Source/drain work function	0.0	eV
Channel doping	-1×10^{15}	cm^{-3}
Source/drain doping	1×10^{20}	cm^{-3}
V_{gs} sweep	0.0–0.7, 14 pts	V
V_{ds} sweep	0.0–0.7, 13 pts	V

Table S3: Default gate insulator parameters (SiO_2 reference).

Parameter	Default value	Unit
Relative permittivity	3.9	—
Electron affinity	0.9	eV
Band gap	9.0	eV
Electron/hole mobility	1×10^{-3}	$\mathrm{m}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$
Electron/hole saturation velocity	2×10^{5}	$\mathrm{m\,s}^{-1}$
Velocity-saturation exponent (electrons)	2.0	—
Velocity-saturation exponent (holes)	1.0	—

The $I_{\mathrm{on}}/I_{\mathrm{off}}$ ratio is used as the primary selector because transistor utility depends on controllable switching rather than raw drive current alone; a highly conductive but leaky channel would be penalised relative to a well-balanced one.

S5 Information Unit Connectivity within the EMOS Framework

Figure S7 captures the core interoperability principle of the framework: each *IU* remains internally independent, but all *IUs* connect through a common property-level schema and consistent data-exchange interfaces. The links in the diagram denote explicit mappings from *IU*-specific field names to canonical property names in the shared property dictionary, allowing semantically equivalent descriptors from different sources to be resolved into a single vocabulary. This mapping layer supports both many-to-one normalisation (multiple source labels mapped to one canonical property) and broad cross-*IU* compatibility, so database outputs can be consumed by predictors and compared with generator outputs without bespoke adapters. In parallel, CIF is used as the common structural representation exchanged across *IU* classes, enabling composition of retrieval, generation, and prediction steps within one workflow graph. The standardised *IU* contracts (`retrieve` for databases, `generate` for generators, and `predict` for predictors) therefore define a stable core, while the mapped property dictionary provides the translation layer needed for extensibility as new modules are integrated. Together, these two design elements convert a heterogeneous ecosystem of tools into a coherent, modular workflow environment that improves reproducibility, benchmarking

consistency, and practical reusability for electronic materials discovery.

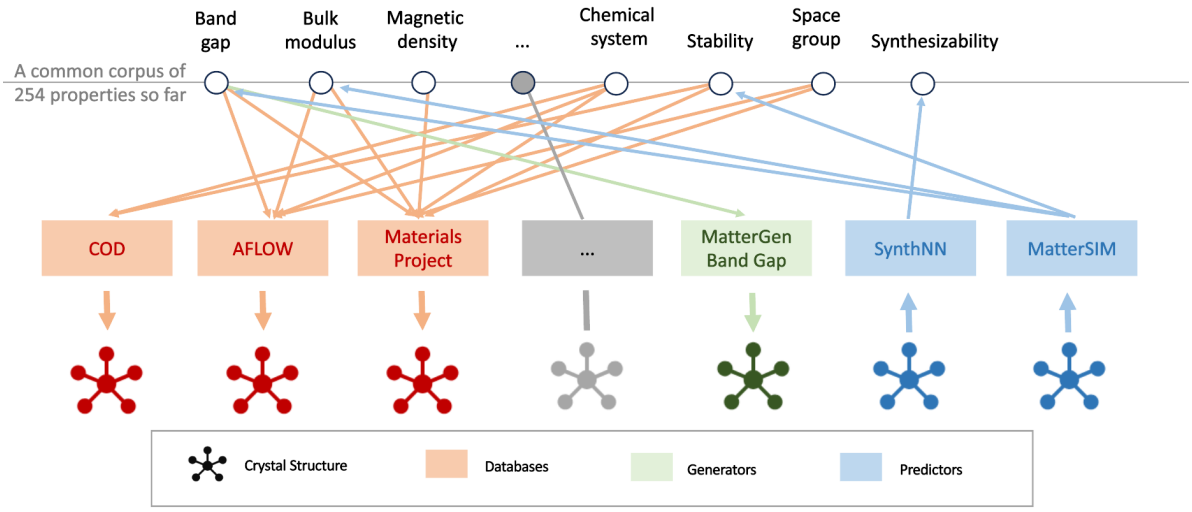


Figure S7: **Connectivity map of *IUs*.** Illustration of database, generator, and predictor modules linked through the shared property dictionary and standardised *IU* interfaces, highlighting canonical-property alignment and cross-module data exchange for workflow composition.

S6 Contribution guidelines

EMOS contributions follow a staged, role-separated workflow that is designed to keep extension effort low while preserving architectural consistency. As shown in Figure S8, the process begins with framework-level design choices (core building blocks, interconnectivity rules, and interface standards), then moves to template and contract definition, and subsequently to backend implementation and tool-assisted UI generation. This separation allows *contributors* to focus on implementing a specific *IU* or *feature* in Python without needing to manually coordinate full-stack integration details, because reusable templates and development tools provide the required scaffolding and enforce compatibility constraints. The resulting modules are then exposed in the user-facing pipeline layer, where *end users* compose *IUs* and higher-level *features* into application-specific workflows. In practice, this contribution model balances flexibility and governance: *administrators* maintain coherence of standards and architecture, *contributors* add functionality with minimal friction, and *end users* immediately benefit from newly integrated components through a consistent interface.

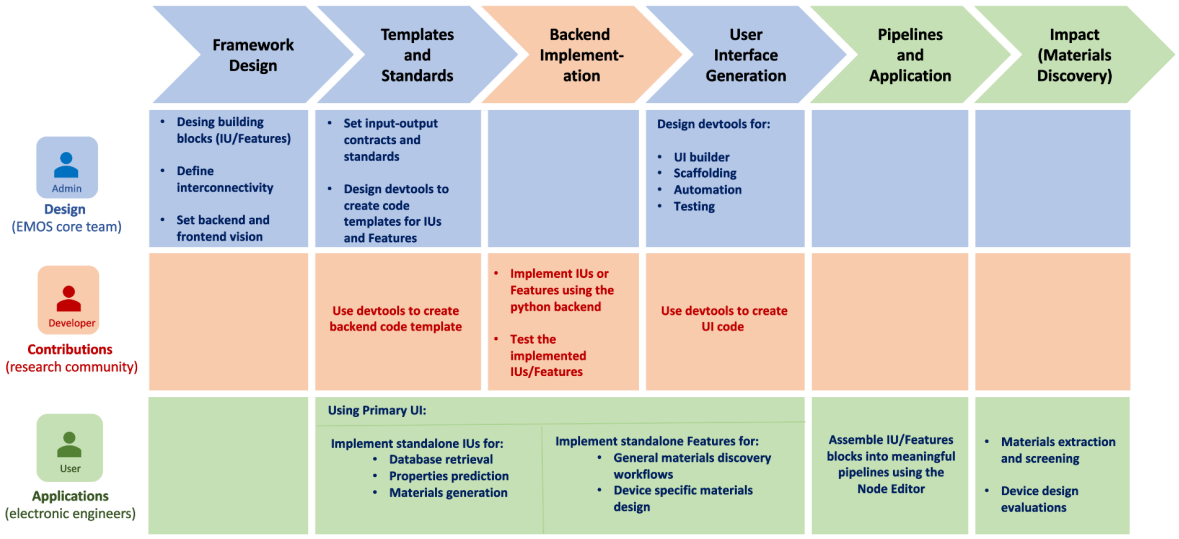


Figure S8: **Role- and stage-structured contribution pipeline in EMOS.** The diagram is organised by three actor lanes (*administrators, contributors* and *end users*) and five development phases, from framework design through template standardisation, backend implementation, and UI generation to user-built pipelines; the rightmost panel summarises downstream application impact.