SUPPORTING INFORMATION

Uncovering structure-property relationships of materials by subgroup discovery

Bryan R. Goldsmith*[,‡,1], Mario Boley[‡,1,2], Jilles Vreeken[2], Matthias Scheffler[1], and Luca M. Ghiringhelli*[,1]

[1]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin, Germany
[2]Max Planck Institute for Informatics, Campus Mitte, 66123 Saarbrücken, Germany

E-mails: *goldsmith@fhi-berlin.mpg.de; ghiringhelli@fhi-berlin.mpg.de

Keywords: big-data analytics, data mining, pattern discovery, machine learning, octet binary semiconductors, gold clusters

Table S1. List of features used in subgroup discovery for the 82 octet binary semiconductors.

IP^B	metric	ionization potential (IP) value of element B (eV)
EA^B	metric	electron affinity (EA) value of element B (eV)
H^B	metric	HOMO (H) value of element B (eV)
L^B	metric	LUMO (L) value of element B (eV)
r_s^B	metric	r_s value of element B (Å)
r_p^B	metric	r_p value of element B (Å)
r_d^B	metric	r_d value of element B (Å)
IP^A	metric	ionization potential (IP) value of element A (eV)
EA^A	metric	electron affinity (EA) value of element A (eV)
H^A	metric	HOMO (H) value of element A (eV)
L^A	metric	LUMO (L) value of element A (eV)
r_s^A	metric	r_s value of element A (Å)
r_p^A	metric	r_p value of element A (Å)
r_d^A	metric	r_d value of element A (Å)
$\lvert IP^A - IP^B \rvert$	metric	derived
$\lvert EA^A - EA^B \rvert$	metric	derived
$\lvert H^A - H^B \rvert$	metric	derived
$\lvert L^A - L^B \rvert$	metric	derived
$\lvert r_s^A - r_s^B \rvert$	metric	derived
$\lvert r_p^A - r_p^B \rvert$	metric	derived
$\lvert r_d^A - r_d^B \rvert$	metric	derived
$\lvert IP^A - IP^B \rvert / IP^A$	metric	derived
$\lvert EA^A - EA^B \rvert / EA^A$	metric	derived
$\lvert H^A - H^B \rvert / H^A$	metric	derived
$\lvert L^A - L^B \rvert / L^A$	metric	derived
$\lvert r_s^A - r_s^B \rvert / r_s^A$	metric	derived

$\lvert r_p^A - r_p^B \rvert / r_p^A$	metric	derived
$\lvert r_d^A - r_d^B \rvert / r_d^A$	metric	derived
EN^A	metric	electronegativity of A (eV)
EN^B	metric	electronegativity of B (eV)
$\lvert H^A - L^A \rvert$	metric	HOMO-LUMO energy gap of A (eV)
$\lvert H^B - L^B \rvert$	metric	HOMO-LUMO energy gap of B (eV)
$\lvert IP^A - EA^A \rvert$	metric	derived
$\lvert IP^B - EA^B \rvert$	metric	derived
$\lvert H^A - L^B \rvert$	metric	derived
$\lvert IP^A - EA^B \rvert$	metric	derived
$\lvert IP^A - EA^A \rvert / r_s^A$	metric	derived
$\lvert IP^B - EA^B \rvert / r_s^A$	metric	derived
$\lvert H^A - L^B \rvert / r_s^A$	metric	derived
$\lvert IP^A - EA^B \rvert / r_s^A$	metric	derived
$\lvert IP^A - EA^A \rvert / r_p^A$	metric	derived
$\lvert IP^B - EA^B \rvert / r_p^A$	metric	derived
$\lvert H^A - L^B \rvert / r_p^A$	metric	derived
$\lvert IP^A - EA^B \rvert / r_p^A$	metric	derived
$\lvert IP^A - EA^A \rvert / r_d^A$	metric	derived
$\lvert IP^B - EA^B \rvert / r_d^A$	metric	derived
$\lvert H^A - L^B \rvert / r_d^A$	metric	derived
$\lvert IP^A - EA^B \rvert / r_d^A$	metric	derived
Δ	metric	energy of RS – energy of ZB
sign(Δ)	categoric	sign of energy difference RS and ZB structures
r_σ	metric	$\lvert r_p^A + r_s^A \rvert - \lvert r_p^B + r_s^B \rvert$ (Å) from *Phys. Rev. Lett.* 1974, *33*, 1095
r_π	metric	$\lvert r_p^A - r_s^A \rvert + \lvert r_p^B - r_s^B \rvert$ (Å) from *Phys. Rev. Lett.* 1974, *33*, 1095
$\lvert r_s^A - r_p^B \rvert \exp(-r_s^A)$	metric	feature 1 from Ghiringhelli *et al*. *Phys. Rev. Lett.* 2015, *114*, 105503
$\lvert IP^B - EA^B \rvert / (r_p^A)^2$	metric	feature 2 from Ghiringhelli *et al*. *Phys. Rev. Lett.*2015, *114*, 105503
$\lvert r_p^B - r_s^B \rvert / \exp(r_d^A)$	metric	feature 3 from Ghiringhelli *et al*. *Phys. Rev. Lett.* 2015, *114*, 105503

Table S2. List of features used in subgroup discovery for the gold clusters.

N	ordinal	number of atoms in the cluster
ΔE	metric	total energy of cluster with respect to its most stable structure at size N (eV)
T	metric	average temperature at which configuration was generated (Kelvin)
0#	ordinal	fraction of atoms with zero bonds
1#	ordinal	fraction of atoms with one bond
2#	ordinal	fraction of atoms with two bonds
3#	ordinal	fraction of atoms with three bonds
4#	ordinal	fraction of atoms with four bonds
5#	ordinal	fraction of atoms with five bonds
6#	ordinal	fraction of atoms with six bonds
7#	ordinal	fraction of atoms with seven bonds
Shape	categoric	3D (nonplanar) and 2D (planar/quasi-planar) based on radius of gyration cut-off
E_{HL}	metric	HOMO-LUMO energy gap (eV)
η	metric	chemical hardness $= [0.5 \times (\varepsilon_{LUMO} - \varepsilon_{HOMO})]$ (eV)
μ	metric	electronic chemical potential $= [0.5 \times (\varepsilon_{LUMO} + \varepsilon_{HOMO})]$ (eV)
$-\varepsilon_{HOMO}$	metric	ionization potential (IP) (eV)
$-\varepsilon_{LUMO}$	metric	electron affinity (EA) (eV)
E_{vdW} / N	metric	many-body dispersion energy per atom (eV per atom)
ΔE_{vdW}	metric	many-body dispersion energy (vdWs) referenced to its maximum at each size (eV)
$\Delta\eta$	metric	chemical hardness referenced to its maximum value at each size (eV)
$\Delta\mu$	metric	electronic chemical potential referenced to its maximum at each size (eV)
$\lvert F \rvert / N$	metric	magnitude of the force per atom for each configuration (eV $Å^{-1}$ $atom^{-1}$)
R_{g0}	metric	radius of gyration of state i that has been normalized by the radius of gyration of the lowest energy planar isomer at size N

Table S3. The radius of gyration cut-offs (R_g^X) used to designate gold clusters (sizes 5-14 atoms) as planar/quasi-planar or nonplanar (compact, three-dimensional), and the radius of gyration of the lowest energy planar isomer at size *N*.[a] The gold cluster structure is considered planar/quasi-planar if $R_g > R_g^X$, otherwise the structure is considered nonplanar.

Size of gold cluster	R_g^X (Å)	R_g(lowest energy planar isomer) (Å)
Au_5	2.12	2.20
Au_6	2.36	2.43
Au_7	2.60	2.68
Au_8	2.80	2.91
Au_9	2.88	2.99
Au_{10}	3.05	3.15
Au_{11}	3.20	3.31
Au_{12}	3.35	3.41
Au_{13}	3.38	3.68
Au_{14}	3.65	3.74

[a] The radius of gyration of the lowest energy planar isomer at size *N* is computed from the fully relaxed structure. R_g^X is chosen by examining the probability distribution of the radius of gyration for all cluster configurations generated by REMD, as well as from analyzing the radius of gyration of the optimized ground state planar/quasi-planar and nonplanar configurations at each size. For sizes where planar/quasi-planar and nonplanar isomer coexistence occurs, a relatively clear gap in the radius of gyration distribution exists. See Figure S4 for examples.

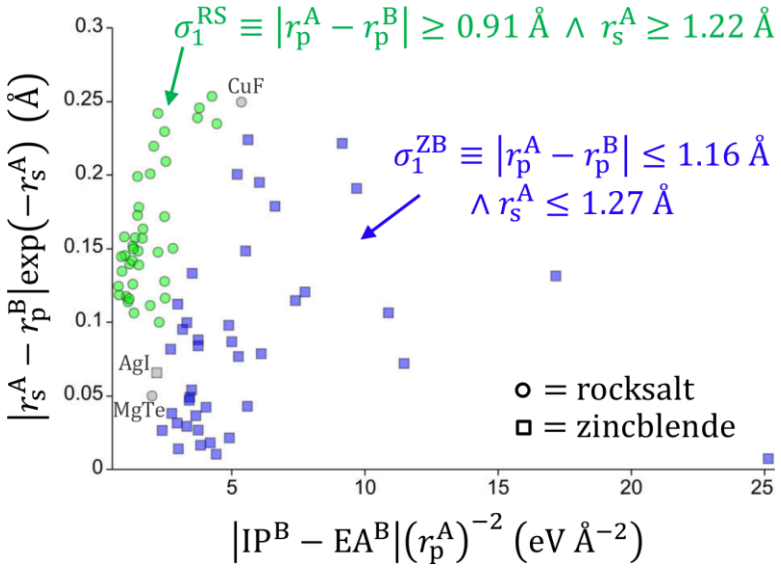

Figure S1. Application of subgroup discovery to the 82 octet binary semiconductors identifies interpretable selectors σ_1^{RS} and σ_1^{ZB} that describe subgroups of the rocksalt (RS) and zincblende (ZB) structures, respectively. Here the axes are chosen to be the two-dimensional descriptor found by Ghiringhelli and coworkers using LASSO+ℓ_0 for visualization purposes only. Green: rocksalt subgroup described by σ_1^{RS}; Blue: zincblende subgroup described by σ_1^{ZB}; Grey: compounds described by neither selector. The circles and squares denote rocksalt and zincblende crystal structures, respectively. 79 of the 82 octet binary semiconductors are described by σ_1^{ZB} and σ_1^{RS}.

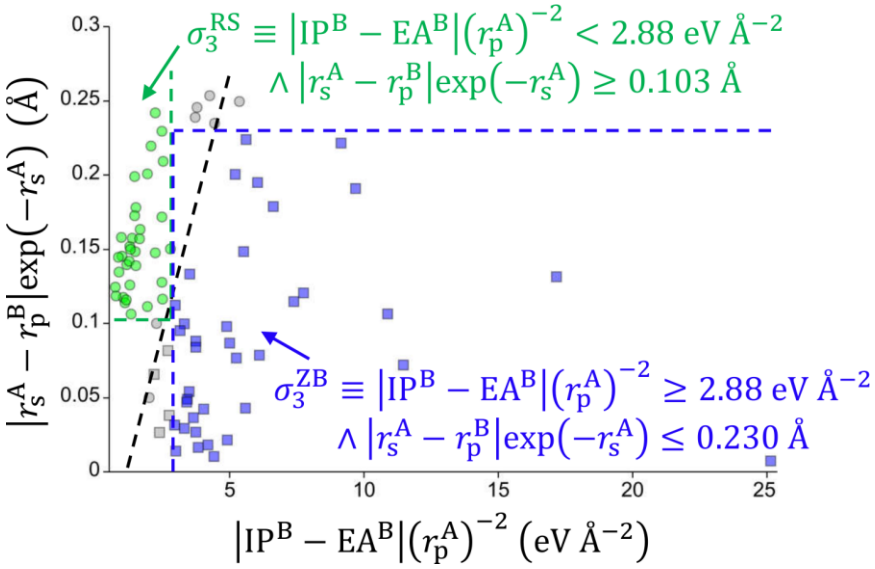

Figure S2. The rocksalt and zincblende subgroups described by selectors consisting of the two-dimensional descriptor found by Ghiringhelli *et al.* using LASSO+ℓ_0. The dashed black line denotes the linear separating hyperplane that the two-dimensional descriptor was originally optimized to describe (by LASSO+ℓ_0). The dashed green and blue lines denote the (non-linear) intersection of axis-parallel hyperplanes that contain the RS and ZB subgroups. Green: rocksalt subgroup described by σ_3^{RS}; Blue: zincblende subgroup described by σ_3^{ZB}; Grey: compounds described by neither selector. The circles and squares denote rocksalt and zincblende crystal structures, respectively. 71 of the 82 octet binary semiconductors are described by σ_3^{ZB} and σ_3^{RS}.

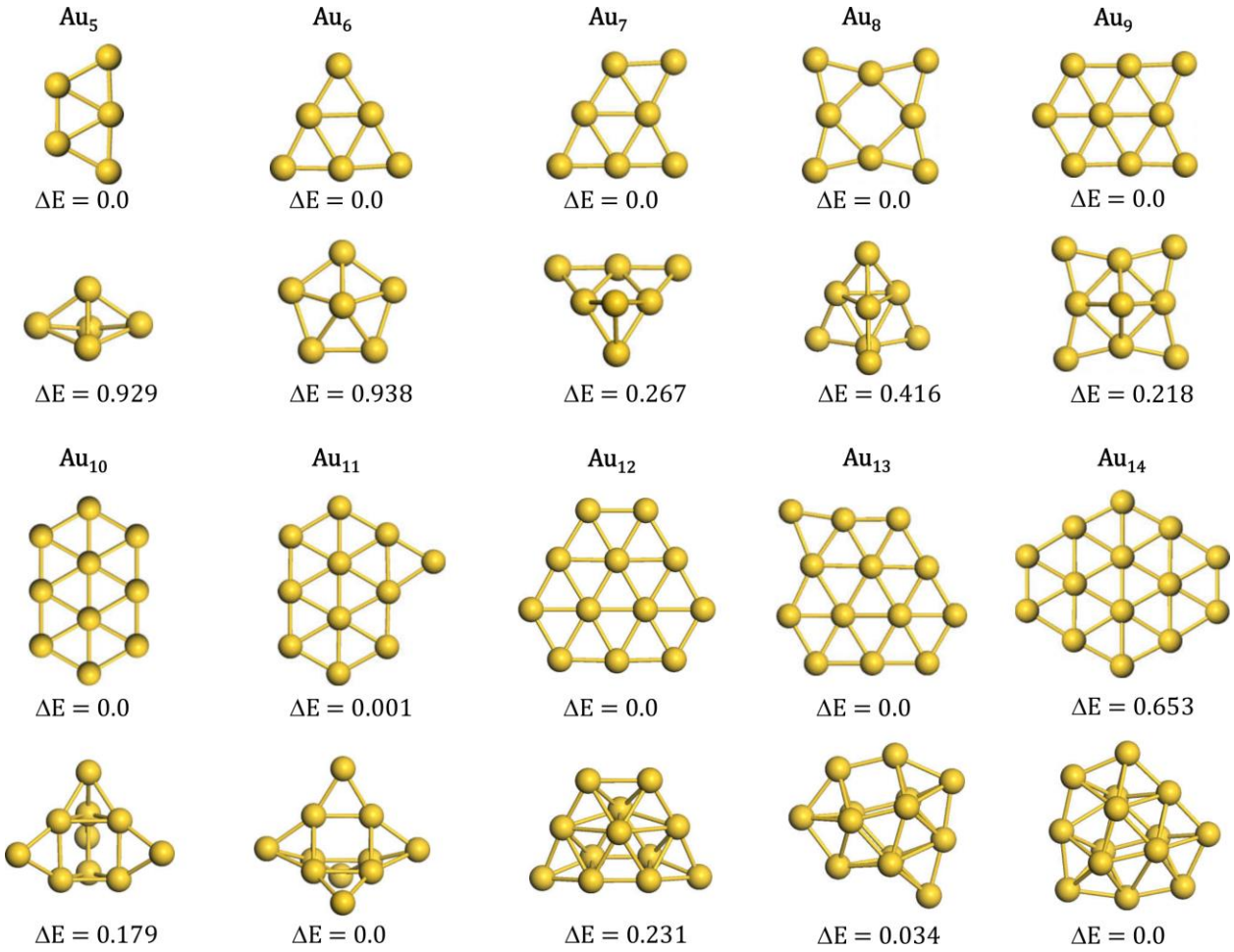

Figure S3. The lowest energy planar and nonplanar gold cluster structures (Au_5-Au_{14}) and their electronic energy differences (in eV). The predicted ground state structure at each size is used as the reference state ($\Delta E = 0.0$). Energies are obtained from fully relaxed structures using PBE+MBD with *tight-tier 2* setting

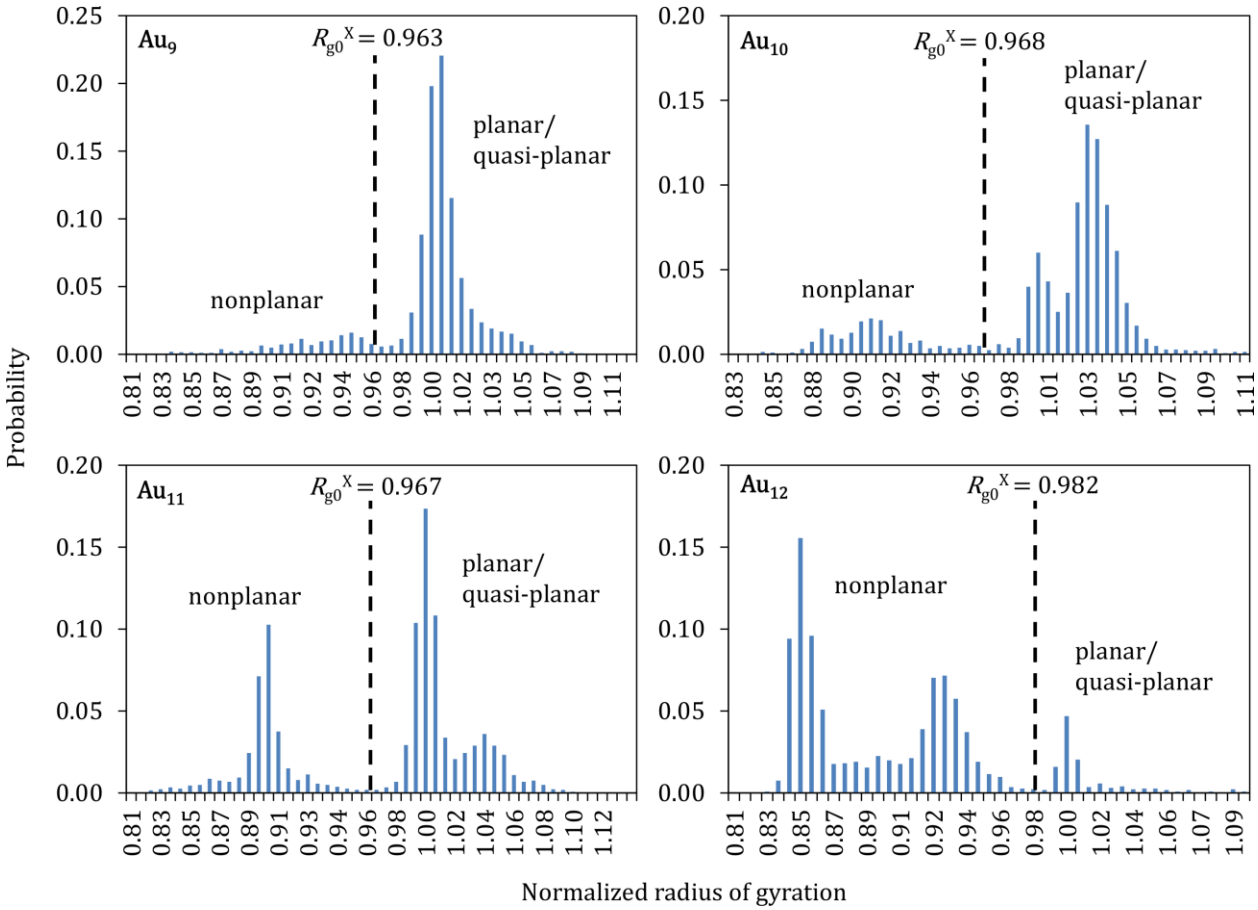

Figure S4. Probability distributions of the normalized radius of gyration (R_{g0}) for Au_9, Au_{10}, Au_{11}, and Au_{12}, which are used to determine the normalized radius of gyration cut-offs (R_{g0}^{X}) that delineate planar/quasi-planar and nonplanar structures. Here $R_{g0}^{X} = R_{g}^{X} \div R_{g}$(lowest energy planar isomer). See Table S3 for the R_{g}^{X} and R_{g}(lowest energy planar isomer) values of Au_5-Au_{14}.