Supporting Information

Predicting Segregation Energy in Single Atom Alloys Using Physics and Machine Learning

Maya Salem, Michael J. Cowan, and Giannis Mpourmpakis*
Department of Chemical Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, United States
*Corresponding author. E-mail: gmpourmp@pitt.edu

1. Derivation of $E_{seg}$ as a function of Cohesive Energy (CE)

Segregation energy ($E_{seg}$; Equation S1) is defined as the difference in energy between a mono-doped metal system when the dopant is on the surface ($E_{dopant,surface}$) and when the dopant is in the bulk ($E_{dopant,bulk}$).

$$E_{seg} = E_{dopant,surface} - E_{dopant,bulk} \ #S1$$

In order to utilize the Bond-Centric Model to calculate $E_{seg}$, we can rewrite Equation S1 as a function of cohesive energy (CE). We first define $CE$ for metallic system $i$ as:

$$CE_i = E_i - \frac{\sum_{j} E_{metal,j}}{n} \ #S2$$

where $E_i$ is the (total electronic) energy of the $n$-atom system and $E_{metal,j}$ is the energy of each metal atom $j$. Applying Equation S2 to the two mono-doped configurations within Equation S1, we find

$$CE_{dopant,surface} = \frac{E_{dopant,surface} - (n-1)E_{metal,host} - E_{metal,dopant}}{n} \ #S3$$

$$CE_{dopant,bulk} = \frac{E_{dopant,bulk} - (n-1)E_{metal,host} - E_{metal,dopant}}{n} \ #S4$$

Note that the summation is simplified to two terms since there are $n-1$ host metal atoms and 1 dopant in each metal system. By subtracting Equation S4 from S3, we achieve
Finally, we can substitute Equation S1 into Equation S5, yielding $E_{\text{seg}}$ as a function of $CE_{\text{dopant, surface}}$ and $CE_{\text{dopant, bulk}}$.

$$E_{\text{seg}} = n(CE_{\text{dopant, surface}} - CE_{\text{dopant, bulk}}) \#S6$$

2. NP Morphology Effects on $E_{\text{seg,BCM}}$

BCM is an empirical model that calculates CE of a NP as a function of the morphology and chemical ordering (i.e. positioning of different metal atoms in the lattice of the NP). To understand the effect of morphology on $E_{\text{seg}}$ of SAAs, we investigate a series of Pt (host) Ni (dopant) SAA NPs.

**Shape**

Elongated pentagonal bipyramid and cuboctahedron at different sizes are examined to study the effect of shape on $E_{\text{seg}}$. An increase in the number of atoms, results in different $E_{\text{seg}}$ of the NP until Pd$_{922}$Ir, as illustrated in Figure S1. This is due to the difference in the coordination number environment present in each shape. This difference becomes apparent in smaller NPs, as a high percentage of atoms are under-coordinated$^1$. At large NPs, $E_{\text{seg}}$ becomes independent of shape.

**Size**

Restricting the shape, an increase in the number of atoms corresponds to an increase in the $E_{\text{seg}}$. This behavior is observed until 923 for CN=9 and 309 for CN=8, 7, and 6, as shown in Figure S1. The slope occurring at small NPs varies depending on the metal combination. The increase in the $E_{\text{seg}}$ with the number of atoms is due to the increase in the first neighbor CN of the dopant. Additionally, $E_{\text{seg}}$ does not change at increased NP size as the CN of the atoms and their neighboring atoms do not change. Similarly, the CN environment in a slab is also constant; hence, large NPs can be used to compute $E_{\text{seg}}$.

**NP Selection**

Since $E_{\text{seg,BCM}}$ is only a function of dopant CN for NPs greater than or equal to 923 atoms (Figure S1), we use the 923-atom NP to compute $E_{\text{seg,BCM}}$ and compare its trends to periodic DFT calculations ($E_{\text{seg,DFT}}$ of periodic surfaces, Figure 1).
Figure S1. (a) $E_{\text{seg,BCM}}$ of PtNi NPs (Pt = host, Ni = dopant) as a function of size and shape. (b) Example NPs to show different sizes and CN sites. (i) 55-atom elongated-pentagonal-bipyramid, (ii) 147-atom cuboctahedron, (iii) 309-atom elongated-pentagonal-bipyramid, and (iv) 923-atom cuboctahedron. At large NP sizes, FCC (111) is represented by CN=9 (green) while FCC (100) is CN=8 (purple), FCC (110) is CN=7 (blue), and FCC (210) is CN=6 (red).
3. List of descriptors used in the Recursive Feature Elimination with Cross-Validation (RFECV) Analysis

Table S1. Descriptors used in the RFECV analyses. Blue highlights correspond to the added descriptors included in the extended feature set. Extended feature set includes all descriptors except the $E_{\text{seg, BCM}}$.

| Descriptor Name                                      | Symbol          | Source          | Number of Features |
|-----------------------------------------------------|-----------------|-----------------|--------------------|
| BCM-calculated segregation energy                   | $E_{\text{seg,BCM}}$ | Calculated herein. | 1                  |
| Lattice constant                                    | $a_i$           | ref$^2$         | 3                  |
| Atomic radius                                       | $r_{i}^{\text{atomic}}$ | ref$^3$        | 3                  |
| Density                                             | $\rho_i$        | ref$^2$         | 3                  |
| Pauling Electronegativity                           | $X_{\text{Pauling},i}$ | ref$^2$       | 3                  |
| Gordy Electronegativity                             | $X_{\text{Gordy},i}$ | ref$^2$       | 3                  |
| First Ionization Potential                          | $IP_i$          | ref$^2$         | 3                  |
| Electron affinity                                   | $EA_i$          | ref$^2$         | 3                  |
| Dipole Polarizability                               | $\alpha_i$      | ref$^2$         | 3                  |
| Bulk cohesive energy                                | $CE_{\text{bulk,i}}$ | ref$^4$       | 3                  |
| Dopant coordination number                          | $CN_{\text{dopant}}$ | Calculated herein. | 1                  |

**Total Size of Initial Feature Set** 29

Bulk Cohesive Energy/Coordination Number

| $CE_{\text{bulk,i}}/CN_{\text{dopant}}$ | Calculated herein. | 3 |

Bulk Cohesive Energy* Coordination Number

| $CE_{\text{bulk,i}} * CN_{\text{dopant}}$ | Calculated herein. | 3 |

Bulk Cohesive Energy/ $\sqrt{\text{Coordination Number}}$

| $CE_{\text{bulk,i}} / \sqrt{CN_{\text{dopant}}}$ | Calculated herein. | 3 |

**Total Size of Extended Feature Set** 38

Symbols with subscript $i$ account for three features: host metal property ($X_{\text{host}}$), dopant metal property ($X_{\text{dopant}}$), and difference between host and dopant metal properties (i.e. $X_{\text{host}} - X_{\text{dopant}}$). For example, $CE_{\text{bulk,i}}; CE_{\text{bulk,host}}, CE_{\text{bulk,dopant}}$, and $\Delta CE_{\text{bulk}} = CE_{\text{bulk,host}} - CE_{\text{bulk,dopant}}$. 
4. Pearson Correlation Table between $E_{\text{seg,BCM}}$ and BCM-inspired features

![Pearson correlation table between $E_{\text{seg,BCM}}$ and BCM-inspired features for FCC metal SAAs.](image)

**Figure S2.** Pearson correlation table between $E_{\text{seg,BCM}}$ and BCM-inspired features for FCC metal SAAs.
5. Predicting DFT $E_{\text{seg}}$ of FCC dopants from our extended feature set.

**Figure S3.** (a) LASSO-based RFECV (b) Pearson’s correlation table based on the five features (labeled in table) used to predict $E_{\text{seg,DFT}}$ of FCC dopants using BCM descriptors.

**Table S2.** Hyperparameters used in the three-feature models ($\Delta CE_{\text{bulk}}/CN_{\text{dopant}}, \chi_{\text{gordy,host}},$ and $r_{\text{dopant}}^{\text{atomic}}$) based on the GridSearchCV results.

| Model | Order | $\alpha$ | $\gamma$ | $\epsilon$ | C |
|-------|-------|----------|----------|-----------|---|
| SVR   | 2nd   | -        | 0.5      | 0.1       | 1 |
| SVR   | 3rd   | -        | 0.5      | 0.2       | 1 |
| KRR   | 2nd   | 0.0015   | 0.03     | -         | - |
| KRR   | 3rd   | 0.03     | -        | 2         | - |
| LASSO | -     | 0.013    | -        | -         | - |
| OLS   | -     | -        | -        | -         | - |

*The values reported are utilized in the models, while the empty blocks containing “-“ indicate that the hyperparameter is not present in the model.*
Figure S4. Parity plot between different regression models using three features ($\Delta C_{\text{bulk}}/C_{\text{dopant}}$, $\chi_{\text{gordy,host}}$, and $r_{\text{dopant}}^{\text{atomic}}$) and $E_{\text{seg,DFT}}$ of FCC dopants.

Table S3. Summary of the three-feature models ($\Delta C_{\text{bulk}}/C_{\text{dopant}}$, $\chi_{\text{gordy,host}}$, $r_{\text{dopant}}^{\text{atomic}}$) examined in Figure S4.

| Model    | Order | MAE (eV) | $R^2$  |
|----------|-------|----------|--------|
| SVR      | 2nd   | 0.459    | 0.55   |
| SVR      | 3rd   | 0.246    | 0.89   |
| KRR      | 2nd   | 0.180    | 0.94   |
| KRR      | 3rd   | 0.128    | 0.98   |
| LASSO    | -     | 0.232    | 0.90   |
| OLS      | -     | 0.231    | 0.90   |
Figure S5. Bootstrapping analysis on different regression models used to predict $E_{\text{seg,DFT}}$ of FCC dopants using three features ($\Delta CE_{\text{bulk}}/CN_{\text{dopant}}, \chi_{\text{gordy,host}},$ and $r_{\text{atomic}}^{\text{dopant}}$).

Figure S6. MAE distribution arising from the second order polynomial KRR test and training set using three features ($\Delta CE_{\text{bulk}}/CN_{\text{dopant}}, \chi_{\text{gordy,host}},$ and $r_{\text{atomic}}^{\text{dopant}}$) based on the bootstrapping analysis.

Table S4. Hyperparameters used in the four-feature models ($\Delta CE_{\text{bulk}}/CN_{\text{dopant}}, \chi_{\text{gordy,host}}, r_{\text{atomic}}^{\text{dopant}}, \Delta E_A$) based on the GridSearchCV results.

| Model | Order | $\alpha$ | $\gamma$ | $\epsilon$ | $C$ |
|-------|-------|----------|----------|-----------|-----|
| SVR   | 2nd   | -        | 0.5      | 0.12      | 1   |
The values reported are utilized in the models, while the empty blocks containing “-“ indicate that the hyperparameter is not present in the model.

**Figure S7.** Parity plot between different regression models using four features ($\Delta CE_{\text{bulk}}/CN_{\text{dopant}}, \chi_{\text{gordy,host}}, r_{\text{dopant}}, \Delta E_{A}$) and $E_{\text{seg,DFT}}$ of FCC dopants.

**Table S5.** Summary of the four-feature models ($\Delta CE_{\text{bulk}}/CN_{\text{dopant}}, \chi_{\text{gordy,host}}, r_{\text{dopant}}, \Delta E_{A}$) examined in Figure S7.
Figure S8. Bootstrapping analysis on different regression models predicting $E_{\text{seg,DFT}}$ using four features ($\Delta CE_{\text{bulk}}/CN_{\text{dopant}}, \chi_{\text{gordy,host}}, r_{\text{dopant}}^{\text{atomic}}, \Delta EA$).

Figure S9. MAE distribution arising from the second order polynomial KRR test and training set using four features ($\Delta CE_{\text{bulk}}/CN_{\text{dopant}}, \chi_{\text{gordy,host}}, r_{\text{dopant}}^{\text{atomic}}, \Delta EA$) based on the bootstrapping analysis.

6. Predicting $E_{\text{seg,DFT}}$ of FCC, BCC, and HCP dopants from our extended feature set.
Figure S10. (a) LASSO-based RFECV (b) Pearson’s correlation table of the five features (labeled in table) selected by RFECV to predict $E_{seg,DFT}$ of FCC, BCC, and HCP dopants.

Table S6. Hyperparameters used in the five-feature models ($\Delta CE_{bulk}/CN$, $\chi_{gordy,host}$, $\Delta EA$, $r_{dopant}^{atomic}$, and $IP_{dopant}$) based on the GridSearchCV results.

| Model  | Order | $\alpha$ | $Y$ | $\epsilon$ | $C$ |
|--------|-------|----------|-----|------------|-----|
| SVR    | 2nd   | 0.5      | 0.15| 0.3        |
| SVR    | 3rd   | -        | 0.5 | 0.2        | 0.3 |
| KRR    | 2nd   | 0.002    | 0.02| -          | -   |
| KRR    | 3rd   | 1.5      | 0.9 | -          | -   |
| LASSO  | -     | 0.0045   | -   | -          | -   |
| OLS    | -     | -        | -   | -          | -   |

The values reported are utilized in the models, while the empty blocks containing “-“ indicate that the hyperparameter is not present in the model.
**Figure S11.** Parity plot between different regression models using five features ($\Delta CE_{\text{bulk/CN}}, \chi_{\text{gordy,host}}, \Delta EA, r_{\text{atomic}},$ and $I_{P_{\text{dopant}}}$) and $E_{\text{seg,DFT}}$ of FCC, BCC, and HCP dopants.

**Figure S12.** Parity plots of $E_{\text{seg,model}}$ (second-order polynomial KRR using five features) vs. $E_{\text{seg,DFT}}$ of FCC metal hosts doped with FCC, BCC, and HCP metal dopants. Data is presented as multiple subplots (where color indicates different metal dopants type) separated by (i) FCC (111), (ii) FCC (100), (iii) FCC (110), and (iv) FCC (210). MAE are determined post-training from the segmented data.
Table S7. Summary of the five-feature models ($\Delta C_{\text{bulk}}/\text{CN}, \chi_{\text{gordy, host}}, \Delta \text{EA}, r_{\text{atomic}}^{\text{dopant}}, \text{and } I_{\text{P dopant}}$) examined in Figure S11.

| Model  | Order | MAE (eV) | $R^2$ |
|--------|-------|----------|-------|
| SVR 2nd | 0.526 | 0.44 |
| SVR 3rd | 0.295 | 0.83 |
| KRR 2nd | 0.220 | 0.91 |
| KRR 3rd | 0.173 | 0.95 |
| LASSO | - | 0.311 | 0.81 |
| OLS | - | 0.310 | 0.81 |

Figure S13. Bootstrapping analysis on different regression models used to predict $E_{\text{seg,DFT}}$ of FCC, BCC, and HCP dopants with five features ($\Delta C_{\text{bulk}}/\text{CN}, \chi_{\text{gordy, host}}, \Delta \text{EA}, r_{\text{atomic}}^{\text{dopant}}, \text{and } I_{\text{P dopant}}$).
Figure S14. MAE distribution arising from the second order polynomial KRR test and training set using five features ($\Delta C_{\text{bulk/CN}}, \chi_{\text{gordy/host}}, \Delta E_A, r_{\text{atomic/dopant}},$ and $I_P_{\text{dopant}}$) based on the bootstrapping analysis.

Figure S15. Parity plots of $E_{\text{seg,model}}$ (second-order polynomial KRR using five features) vs. $E_{\text{seg,DFT}}$ of FCC metal hosts doped with FCC, BCC, and HCP metal dopants. Data is presented as (a) a single plot of all the results (where color represents host metal) and (b) multiple subplots (where color indicates surface facet) separated by (i) FCC, (ii) BCC, and (iii) HCP dopants. MAE for the model (a) is calculated from LOOCV while the subplot MAEs (b) are determined post-training from the segmented data.
Figure S16. The MAE of the predicted $E_{\text{seg}}$ on our extended dataset results in Figure S15 separated by (a) the different metal hosts and (b) the different dopant crystal structure Ni-host metal systems.

Table S8. The electronic energies of the homo-atomic metals used for the $\gamma$ calculations utilized in the BCM.

| Metals | Multiplicity | Electronic Energy (eV) |
|--------|--------------|------------------------|
| AgAg   | triplet      | -7998.920172           |
| PtPt   | triplet      | -6495.734119           |
| RhRh   | septet       | -6015.218075           |
| NiNi   | quintet      | -82060.744423          |
| IrIr   | quintet      | -5678.664159           |
| CuCu   | singlet      | -89254.589349          |
| AuAu   | triplet      | -7389.490208           |
| PdPd   | quintet      | -6959.292543           |

Table S9. The electronic energies of the hetero-atomic metals used for the $\gamma$ calculations utilized in the BCM.

| Metals | Multiplicity | Electronic Energy (eV) |
|--------|--------------|------------------------|
| PtAg   | doublet      | -7246.973877           |
| PtPd   | triplet      | -6727.582820           |
| PtAu   | quartet      | -6942.242521           |
| PtNi   | triplet      | -44278.570155          |
| PtCu   | quartet      | -47875.219856          |
| PtIr   | sextet       | -6087.108317           |
| RhAg   | triplet      | -7006.514762           |
| RhPd   | doublet      | -6486.695399           |
| RhAu   | triplet      | -6702.113382           |
| Metal  | Multiplicity | Electronic Energy (eV) |
|--------|--------------|------------------------|
| RhNi   | sextet       | -44038.338357          |
| RhCu   | quintet      | -47634.522118          |
| RhIr   | septet       | -5847.165961           |
| PdNi   | quintet      | -44510.623011          |
| PdIr   | quartet      | -6318.829729           |
| IrAg   | triplet      | -6837.839701           |
| IrAu   | quintet      | -6533.198381           |
| IrNi   | sextet       | -43869.832941          |
| IrCu   | quintet      | -47466.020807          |

Table S10. The electronic energies of the single metal atoms used for the $\gamma$ calculations utilized in the BCM.

### References

1. Yan, Z.; Taylor, M. G.; Mascareno, A.; Mpourmpakis, G. Size-, Shape-, and Composition-Dependent Model for Metal Nanoparticle Stability Prediction. *Nano Letters* **2018**, *18*, 2696-2704.

2. Mentel, L. {Mendelev} -- a Python Resource for Properties of Chemical Elements, Ions and Isotopes. 2014. Available at: https://github.com/lmmentel/mendelev.

3. Clementi, E.; Raimondi, D. L.; Reinhardt, W. P. Atomic Screening Constants from Scf Functions. Ii. Atoms with 37 to 86 Electrons. *The Journal of Chemical Physics* **1967**, *47*, 1300-1307.

4. Jana, S.; Sharma, K.; Samal, P. Assessing the Performance of the Recent Non-Empirical Semilocal Density Functionals on Describing the Lattice Constants, Bulk Moduli and Cohesive Energies of Alkali, Alkaline-Earth, and Transition Metals. *The Journal of Chemical Physics* **2018**, *149*, 164703.