Report on Project 2: Alloy cluster expans ions

Erik Karlsson Öhman, Hampus Hansen

August 5, 2025

Task 1

The Cu concentration as a function of the unstandardized mixing energy is presented in Figure 1.

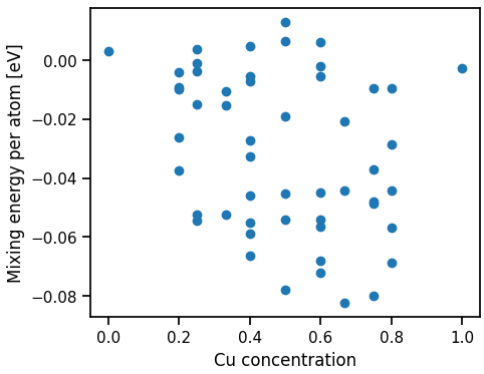


Figure 1: Cu concentration of the structures in the database, plotted against mixing energy per atom (eV).

The mixing energy and the cluster vectors are standardized with the `StandardScaler` from the `sklearn.preprocessing` library, i.e. the mean of the data is set to $\mu = 0$ and the standard deviation is set to $\sigma = 1$ in the following way: $X_{\text{std}} = \frac{X - \bar{X}}{\sqrt{\text{Var}(X)}}$. Standardizing data is good practice since it ensures that the data is all on the same scale and centered around zero. This is especially important when MCMC sampling, as it allows the walkers to explore the high dimensional parameter space effectively without needing to rescale the step size depending on the direction the walkers take.

Task 2

To fit the ECIs using OLS and Ridge regression the `Sklearn`-library functions `linear_regressor()` and `Ridge()` is used. The former corresponds to minimizing the loss function $\mathcal{L}_{\text{OLS}} = ||E - XJ||^2$, or equivalently solving

$$J_{\text{OLS}} = (X^T X)^{-1} X^T E. \tag{1}$$

The latter includes the regularization term in the loss function $\mathcal{L}_{\text{Ridge}} = ||E - XJ||^2 + \alpha||J||^2$, equivalent to solving

$$J_{\text{Ridge}} = (X^T X + \alpha I)^{-1} X^T E. \tag{2}$$

The optimal value of the hyperparameter α was found by cross validating with three folds, and minimizing the cross-validation root-mean-square-error (CV-RMSE). Three folds allowed the validation data set to be large enough to find an optimal value, $\alpha = 0.3416$, as seen in Figure 2.

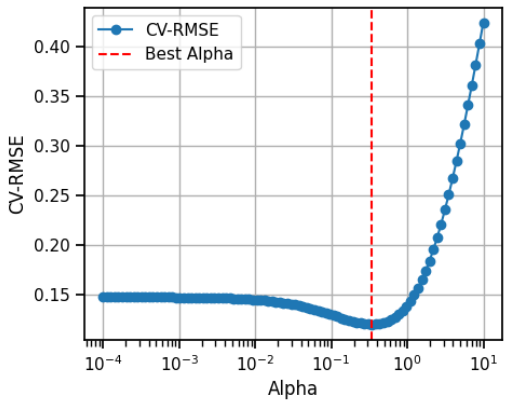


Figure 2: CV-RMSE as a function of α, where the red dashed line markes the optimal α that minimizes CV-RMSE.

The fitted ECIs for the two methods can be found in Figure 3, where we note that most ECIs reduce in size, some more drastically than others, while a few increase in size. This is attributable to the penalization of larger ECIs due to the regularization term, effectively negating overfitting. The corresponding CV-RMSE scores are 0.1472 for OLS and 0.1198 (Ridge), indicating that the Ridge regression is preferred over OLS.

Task 3

The Bayesian Ridge regression method (BRR) is here characterized by the regression matrix Λ, with non-identical diagonal entries λ_α and zero off-diagonal entries $\lambda_{\alpha,\beta} = 0$ [1]. The optimal ECIs are thus found by

$$J_{\text{BRR}} = (X^T X + \Lambda)^{-1} X^T E. \tag{3}$$

The linear regularization scheme

$$\lambda_\alpha(n, r, \gamma) = \gamma_1 r + \gamma_2 n \tag{4}$$

allows parametrization of the entries with hyperparameters γ_1 and γ_2, where n and r is the number of sites and radius of orbit α respectively. The minimization of the CV-RMSE with regards to the hyperparameters γ is done with `scipy`-function `minimize()` and the method L-BFGS-B. The optimal gamma is $\gamma_1 = 0.1439\,\text{Å}^{-1}$, $\gamma_2 = 0.02423$.

The comparison between the three methods can be found in Figure 3. For many ECIs, the Bayesian ridge is more similar to OLS than Ridge. The CV-RMSE is however lower for BRR than the other two methods, at 0.1052. This might indicate that the Ridge model is underfitted and OLS is overfitted, while BRR is better balanced between under- and overfitting. The BRR model is thus preferred to the previous two models.

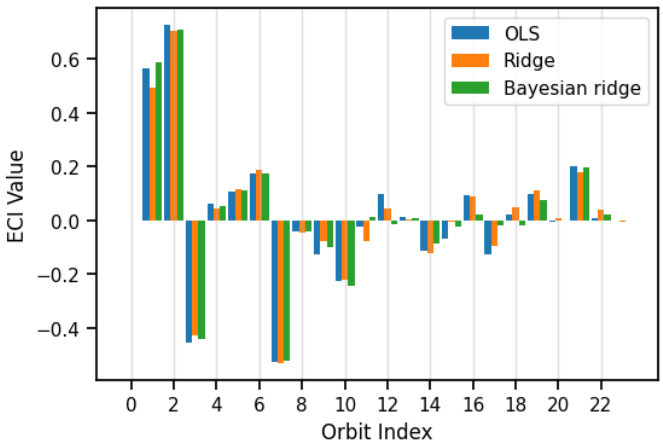


Figure 3: Size of the ECIs for the OLS-, Ridge- and BRR-model in blue, orange and green respectivly.

An individual parameter λ_α for each orbit allows encoding prior beliefs about the system by selectively penalizing different orbits and radii. Thus we can encode our physical intuition - that smaller radii and lower order should contribute more to the energy, as seen in Equation (4). Ridge regression, on the other hand, assumes all parameters are equally likely to be penalized, not allowing this encoding of our physical intuition.

Task 4

The prior for the ECIs was set to a gaussian

$$P(\boldsymbol{J}) = \frac{1}{(2\pi\alpha^2)^{N_p/2}} \exp\left(-||\boldsymbol{J}||^2/2\alpha^2\right) \tag{5}$$

where N_p is the number of ECIs. For α and σ the priors were set to inverse gamma distributions with cutoffs at 0.02 and 2 for σ, and cutoffs at 0.05 and 0.5 for α. The shape and scale parameters of the two distributions were both set to 1. For the likelihood function, a gaussian distribution was used.

For the MCMC sampling a total of $78 = 3N_{dim}$ walkers were used. Each walker sampled 40000 steps each, whereof the initial 15000 steps was used as a burn-in an therefore discarded. Furthermore, to ensure that the samples are uncorrelated, a thinning of 50 was used, such that the autocorrelation time, τ, fulfilled $\tau > N_{\text{samples}}/50$. The resulting posterior distributions are shown in Figure 4 and 5.

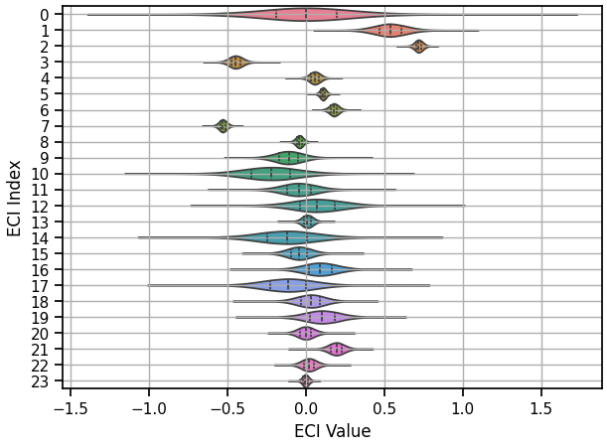


Figure 4: Posterior distributions for all ECIs. The dashed lines on each distribution are the quartiles.

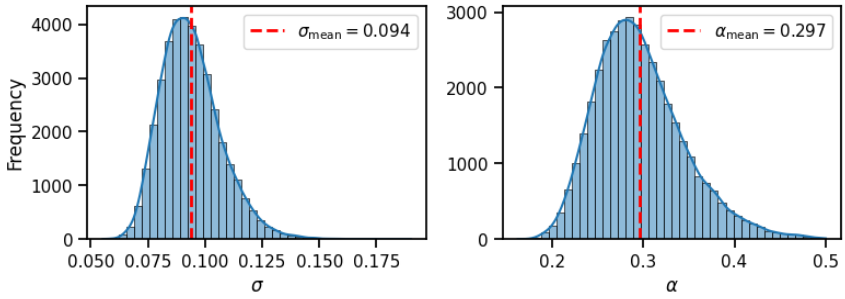


Figure 5: Posterior distributions for σ and α. The red dashed lines marks the mean value of the distributions.

There are 10 parameters where zero lies within the 25th and 75th percentile (specifically for ECI index $\{0, 11, 12, 13, 14, 15, 18, 20, 22, 23\}$). These parameters could be regarded as unnecessary compared to the other 14 parameters, which are more distinguished from zero.

If an unphysical prior was set, for instance favoring higher order clusters, we expect the posterior distributions of the corresponding ECIs grow. This would drastically overestimate the contribution to the total energy to larger order clusters, and simultaneously underestimate the contributions from the smaller order clusters.

Task 5

Performing a parameter sweep over $\lambda_{\text{threshold}}$ we see that $\lambda_{\text{threshold}} = 9331.90$ minimizes the cross-validation error, where we once again use k-folds cross validation with 3 folds as in previous tasks, see Figure 6.

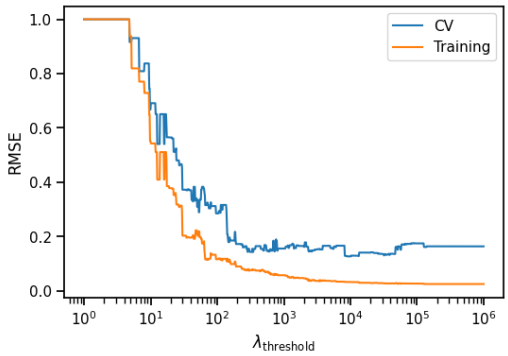
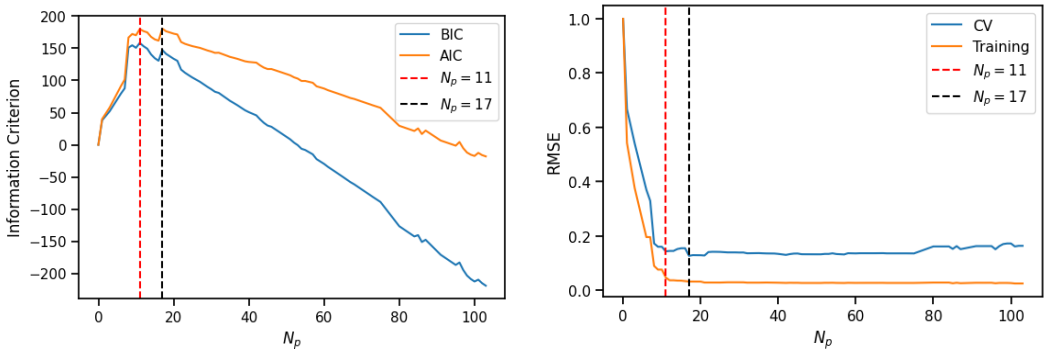


Figure 6: Root mean squared error of cross validation and training error, as a function of the $\lambda_{\text{threshold}}$ parameter.

Analyzing AIC and BIC scores we find that the number of nonzero parameters that maximize the BIC and AIC scores is either $N_p = 11$ or $N_p = 17$, with a slight preference for $N_p = 11$ according to BIC and $N_p = 17$ according to AIC, as seen in Figure 7(a). Analyzing the cross-validation and training error as a function of the number of non-zero parameters, we see that the cross-validation error is minimized for a model with $N_p = 17$, as seen in Figure 7(b). Even though $N_p = 17$ minimizes the cross-validation error and maximizes the AIC score, $N_p = 11$ might be a more suitable choice. The error increases only a negligible amount from the minimum at $N_p = 17$ to $N_p = 11$ and $N_p = 11$ maximizes the BIC score. Furthermore the corresponding AIC score at $N_p = 11$ is only somewhat lower than the global maxima at $N_p = 17$. When choosing $N_p = 11$, the resulting $\lambda_{\text{threshold}}$-value becomes $\lambda_{\text{threshold}} = 286.06$. This value is subsequently used in task 6.



(a) AIC/BIC as a function of non-zero parameters.

(b) Root mean squared error for training and cross validation, as a function of non-zero parameters.

Figure 7: Information criterias and RMSE as a function of non-zero parameters.

When visualizing the value of the ECIs for $\lambda_{\text{threshold}} = 286.06$, we see in Figure 8 that the ECIs are noticeably sparse, i.e. most of the ECIs are set to zero. This is expected since ARDR performs feature selection.

Task 6

Using the models from previous tasks, we see that OLS, Ridge regression and ARDR all predict AuCu as being the ground state structure, with predicted energies in the range

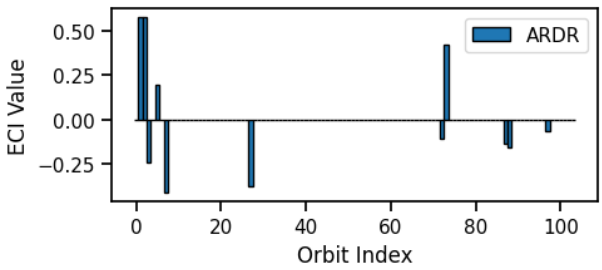


Figure 8: ECIs for ARDR.

-0.105 to $-0.108\,\mathrm{eV}$ as seen in Table 1. The Bayesian Ridge regression, however, predicts AuCu3 as being the most likely ground state structure, with an energy of $-0.114\,\mathrm{eV}$.

Table 1: Ground state candidates that minimize the predicted ground state energy, and the corresponding ground state energies.

	OLS	Ridge	Bayesian Ridge	ARDR
GS candidate	AuCu	AuCu	$AuCu_3$	AuCu
E_{GS} (eV)	-0.108	-0.107	-0.114	-0.105

From the MCMC sampling, the $AuCu_3$ structure was identified as the ground state in 46% of the sampled models, AuCu in 31%, and Au_3Cu in 23%. When analyzing the distribution of the ground state energy obtained from the MCMC sampling we see in Figure 9 that the predictions from OLS and Ridge regression lay close to the peak of the distribution. The predictions from Bayesian Ridge regression are noticeably lower than the peak, and the ARDR model predicts the highest energy. The latter is however relatively sensitive to the number of parameters chosen. If more parameters are included, it might improve predictions.
Since BRR and a full Bayesian analysis is conditioned on physical knowledge about the system we can conclude that these models are preferred to the automatic feature selection models. They also predict a different ground state than the other models, indicating that our prior physical knowledge is key in order to construct a model with accurate predictions. Furthermore, BRR is not as costly as performing a full Bayesian analysis - so in this case, BRR is the preferred method.

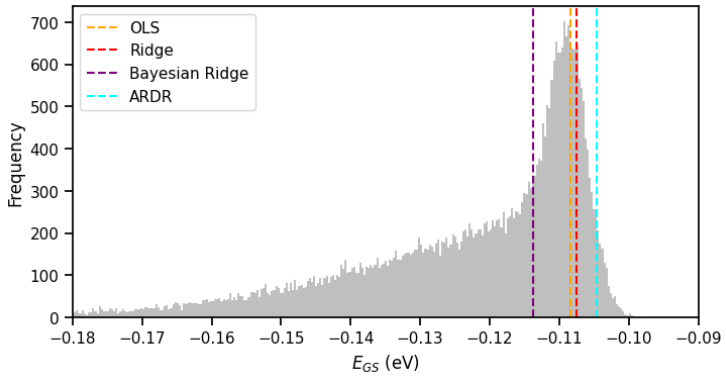


Figure 9: Distribution of the ground state energy, E_{GS}, obtained from MCMC sampling.

References

[1] P. Erhart, *Advanced Simulation and Machine Learning: Project 2a*. Gothenburg, Sweden: Chalmers University of Technology, Dec. 3, 2024.