Supplementary Information to Quantitative NCI index: Defining the link between NCI index and interaction energy

Katarzyna J. Zator*,† and Julia Contreras-García*,†,‡

†*Laboratoire de Chimie Théorique (LCT), Sorbonne Université, UMR 7616, 4 place Jussieu, 75005, Paris*

‡*CNRS*

E-mail: katarzyna.zator@sorbonne-universite.fr; julia.contreras_garcia@sorbonne-universite.fr

Phone: +33 1 44 27 38 79

1 Symbolic Regression

1.1 Internal regression parameters

The following are PySR parameters which were quickly optimized to prevent under- and over-fitting, and excessively long calculations. Train-test split: test_size=0.33, random_state=42) Regression parameters: binary_operators=["mult", "plus"], unary_operators=["sqrt","root3(x) = cbrt(x)"], extra_sympy_mappings="root3": lambda x: x**(1/3), procs=21, verbosity=False, progress=False, turbo=True, deterministic=False, niterations=500, maxsize=10, maxdepth=20, populations=50, constraints='sqrt': (1,1), 'root3': (1,1), 'mult': (3, 3), nested_constraints = "root3": "root3": 0, "sqrt": "root3": 0, "root3": "sqrt": 0, complexity_of_variables=2,

For promolecular densities, all 432 combinations of parameters were studied, but for DFT

densities, only $\rho_c = 0.07$ was considered due to previous studies.

1.2 Variation in equations for an example set of parameters

As the PySR algorithm employs the genetic algorithm, it is inherently non-deterministic and the resulting equations it produces, vary every time the model is fitted. Nevertheless, due to the breadth of the search employed by the PySR package prevents the variability of the equations from being too large as the search is extensive enough to find the optimum result for the given constraints. To showcase the forms of equations obtained, below we list equations obtained from repeated PySR fits. The NCIPLOT parameters used are $s_c = 1.0$ and $\lambda_{small} = 0.02$ for specified $\lambda_{large} = 0.2$, $\gamma_{ref} = 0.85$ and $\rho_c = 0.07$ for the HB375 dataset.

$$E_{Hydrogen\_bond}(\rho) = -(1.805 \times 10^3 \sqrt{I_{3,Hydrogen\_bond}} + 2.928 \times 10^2 \times I_{3/2,van\_der\_Waals}) \quad (1)$$

$$E_{Hydrogen\_bond}(\rho) = -(1.765 \times 10^3 \sqrt{I_{3,Hydrogen\_bond}} + 2.852 \times 10^2 \times I_{2,van\_der\_Waals}) \quad (2)$$

$$E_{Hydrogen\_bond}(\rho) = -(1.200 \times 10^3 I_{5/3,Hydrogen\_bond} + 7.457 \times 10^2 \times \sqrt{I_{5/2,van\_der\_Waals}}) \quad (3)$$

with respective MAE (in kJ/mol): 3.03, 3.05, 3.40, and R^2 coefficients: 0.86, 0.85, 0.83. Should we additionally allow the train and test dataset compositions to vary, the equations are still similar:

$$E_{Hydrogen\_bond}(\rho) = -(1.809 \times 10^3 \sqrt{I_{3,Hydrogen\_bond}} + 2.9250 \times 10^2 \times I_{3/2,van\_der\_Waals}) \quad (4)$$

$$E_{Hydrogen\_bond}(\rho) = -(1.754 \times 10^3 \sqrt{I_{3,Hydrogen\_bond}} + 6.610 \times 10^2 \times I_{5/3,van\_der\_Waals}) \quad (5)$$

$$E_{Hydrogen\_bond}(\rho) = -(1.768 \times 10^3 \sqrt{I_{3,Hydrogen\_bond}} + 2.841 \times 10^3 \times I_{5/2,van\_der\_Waals}) \quad (6)$$

with respective MAE (in kJ/mol): 3.26, 3.12, 3.06, and R^2 coefficients: 0.82, 0.88, 0.85.

Additionally, we tested the effect of NCI indices with $n = 0$ (which represent volume rather than charge) which could have shown meagre improvement in D1200 fits only (2.11 vs. 2.30 kJ/mol) and no effect for the HB375 dataset. The combined equation (again excluding the dispersion term in the HB375 equation) also showed a slight improvement (2.53 vs. 2.63 kJ/mol). However, this has not shown to be transferrable, giving a much worsened 5.2 kJ/mol result for the S66 lateral test versus the original 4.9 kJ/mol; therefore, we decided against utilizing the $n = 0$ terms.

1.3 Parametrisation of HB375 dataset with promolecular densities

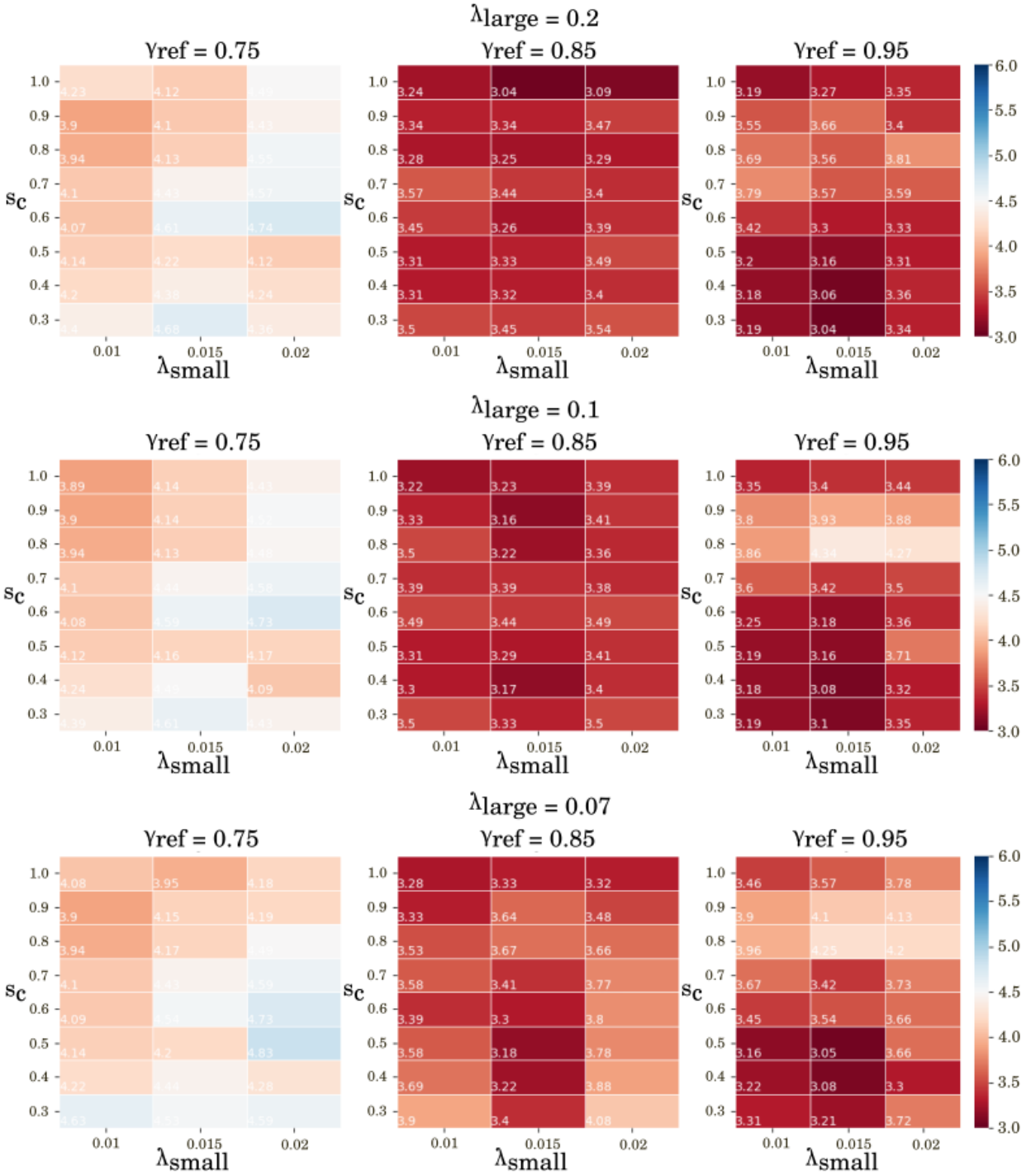


Figure 1: Heat map of mean absolute error (MAE) for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.05$ for the HB375 dataset. All underlying NCI indices calculations were carried out using the promolecular approach and errors are in kJ/mol.

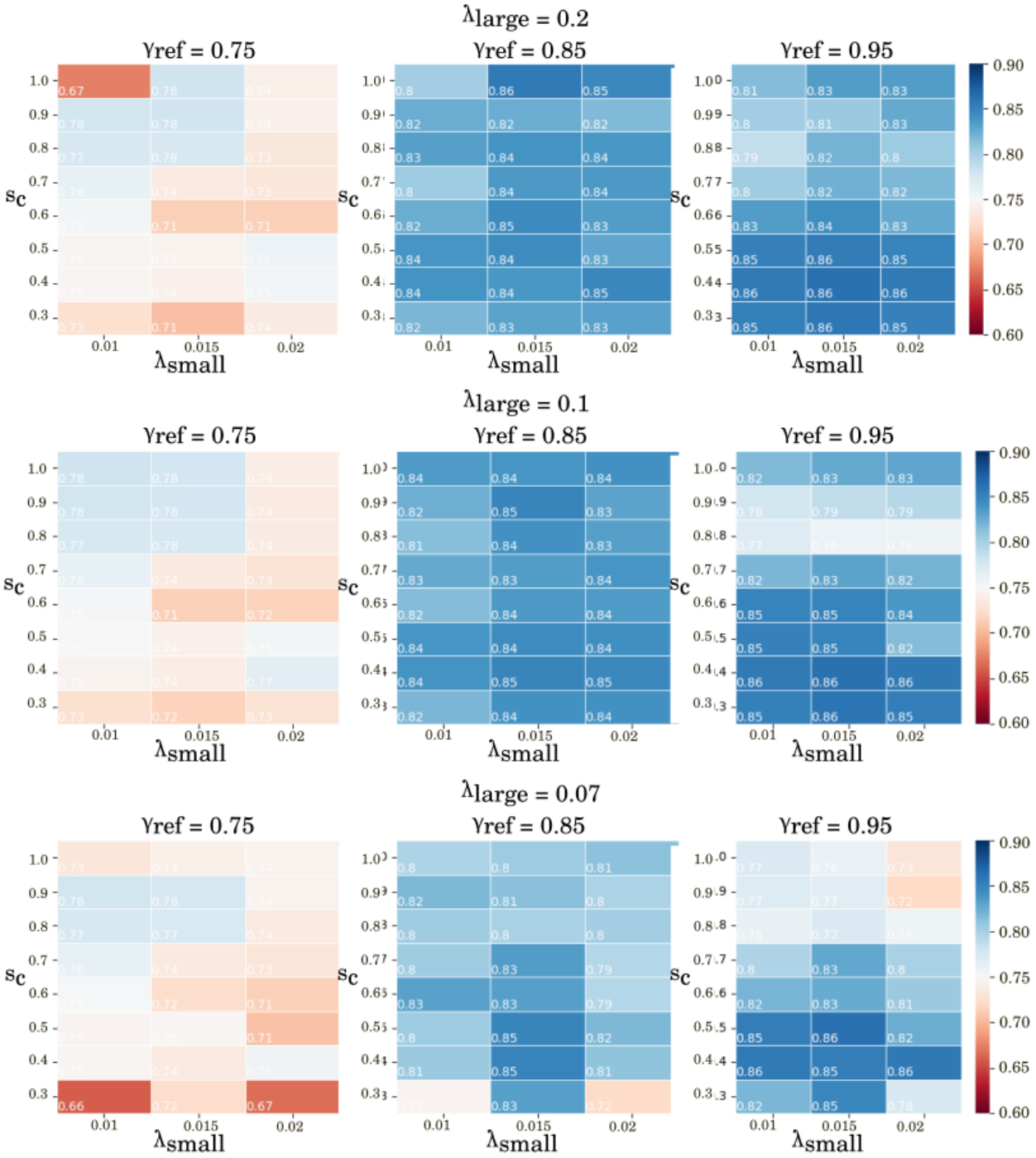


Figure 2: Heat map of R^2 for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.05$ for the HB375 dataset. All underlying NCI indices calculations were carried out using the promolecular approach.

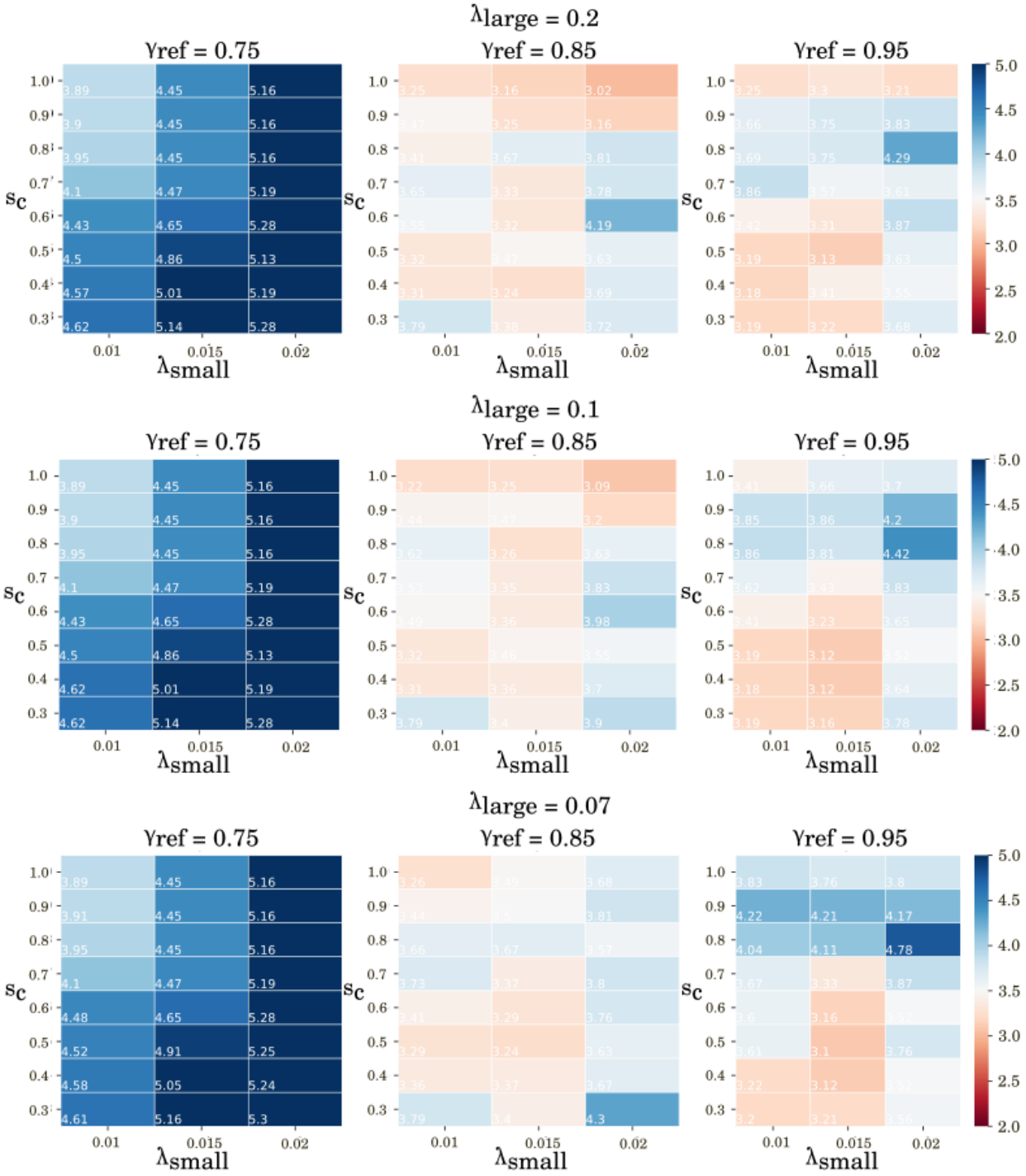


Figure 3: Heat map of mean absolute error (MAE) for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the HB375 dataset. All underlying NCI indices calculations were carried out using the promolecular approach and errors are in kJ/mol.

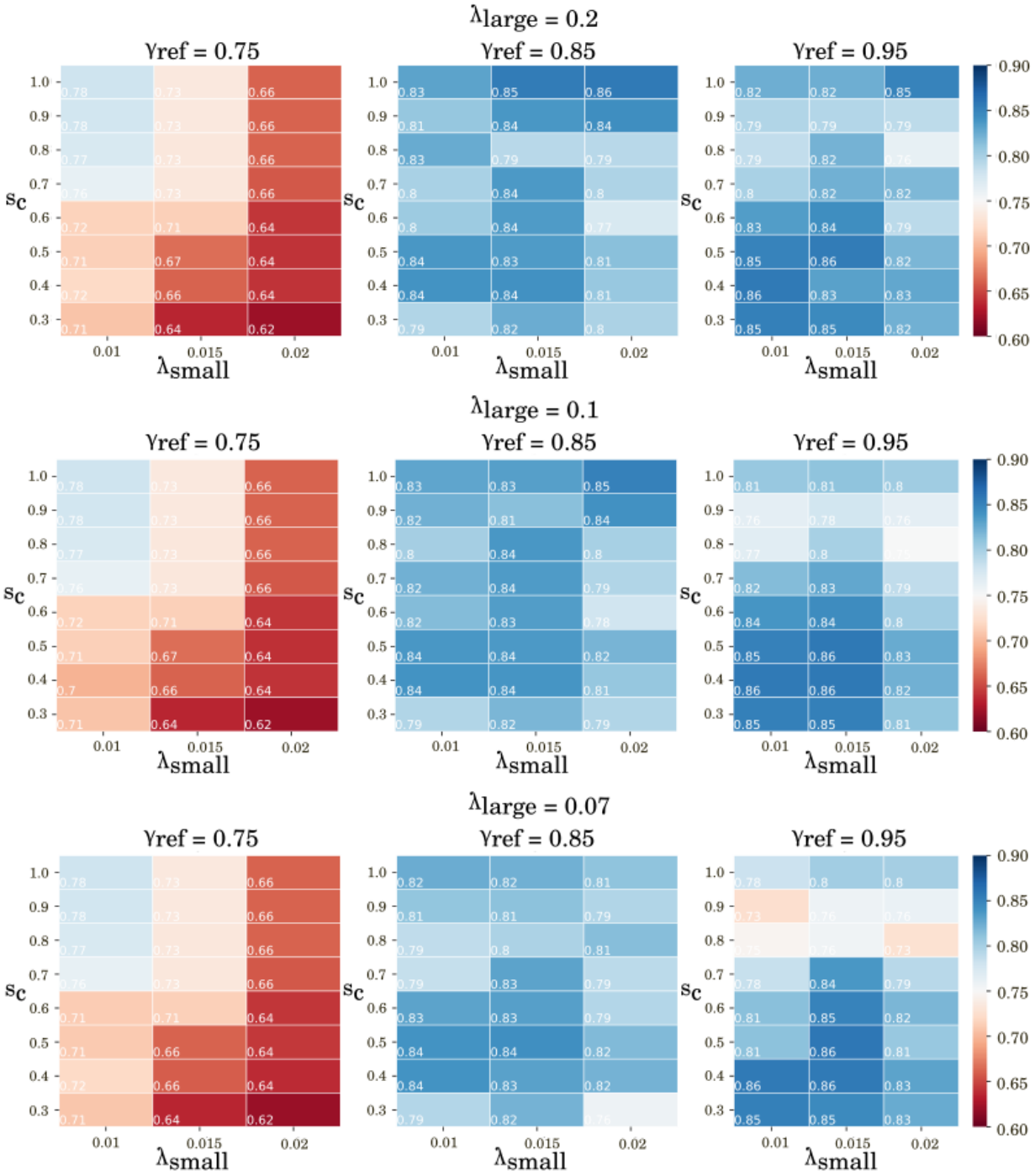


Figure 4: Heat map of R^2 for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the HB375 dataset. All underlying NCI indices calculations were carried out using the promolecular approach.

1.4 Parametrisation of D1200 dataset with promolecular densities

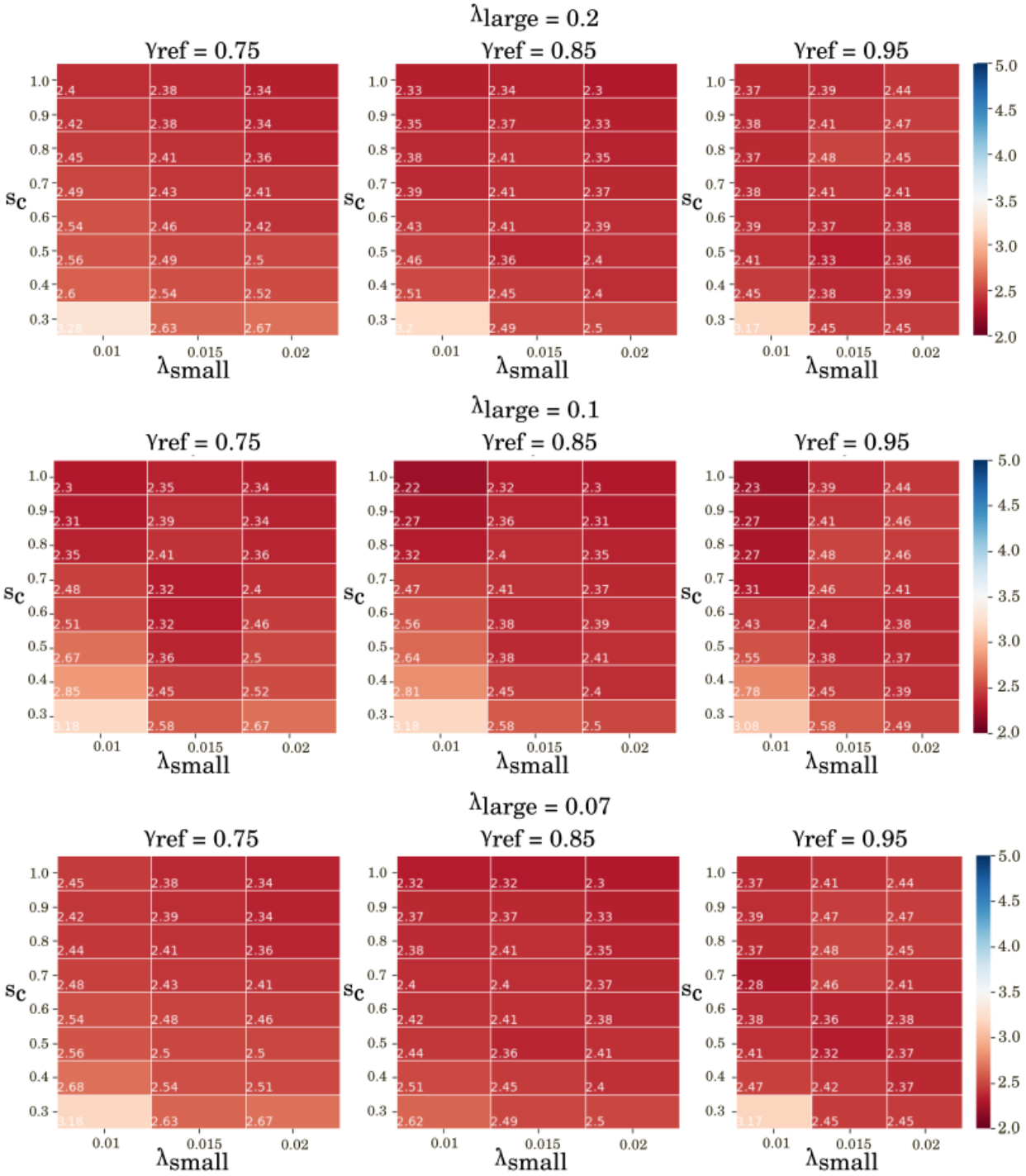


Figure 5: Heat map of mean absolute error (MAE) for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.05$ for the D1200 dataset. All underlying NCI indices calculations were carried out using the promolecular approach and errors are in kJ/mol.

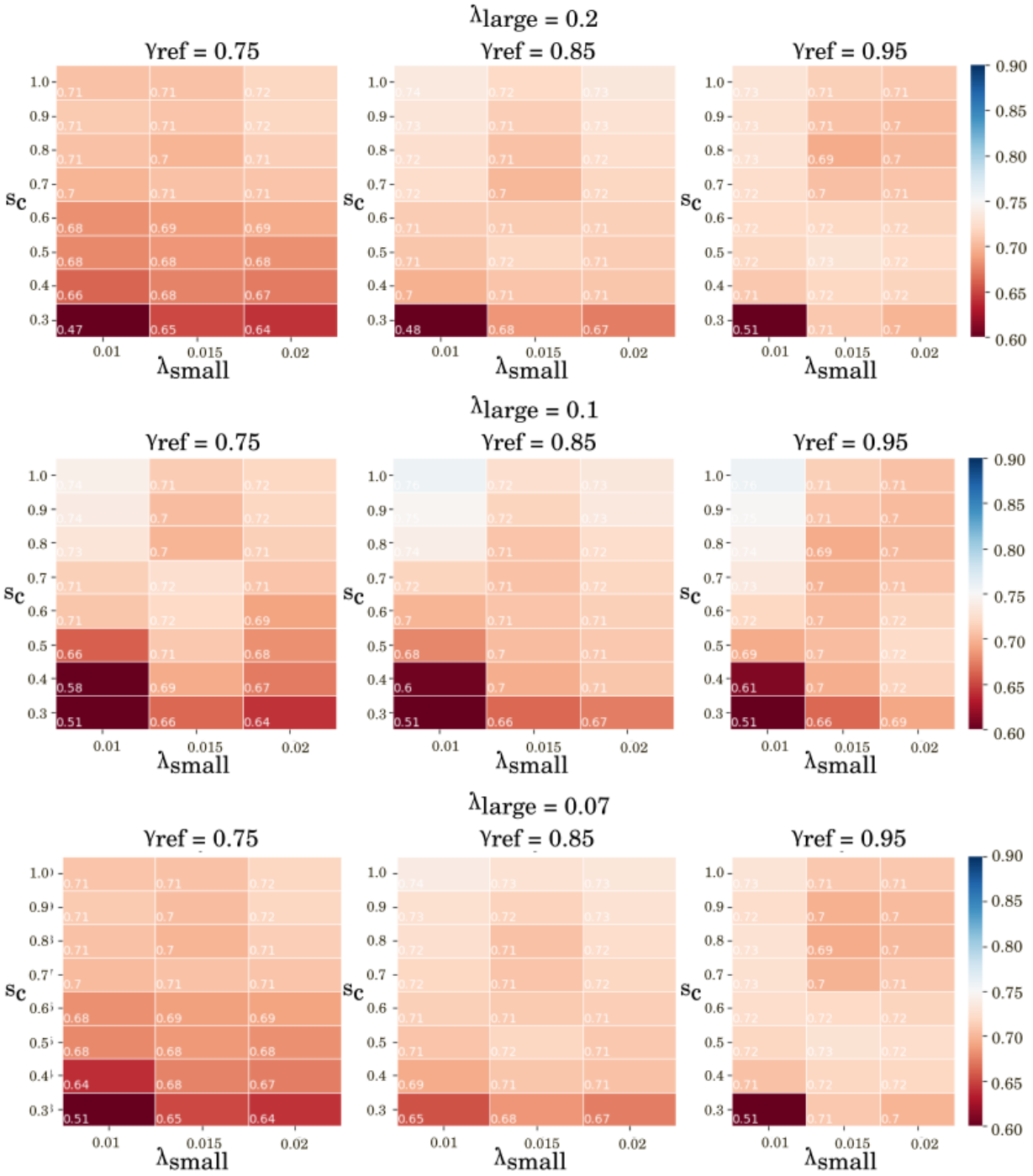


Figure 6: Heat map of R^2 for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.05$ for the D1200 dataset. All underlying NCI indices calculations were carried out using the promolecular approach.

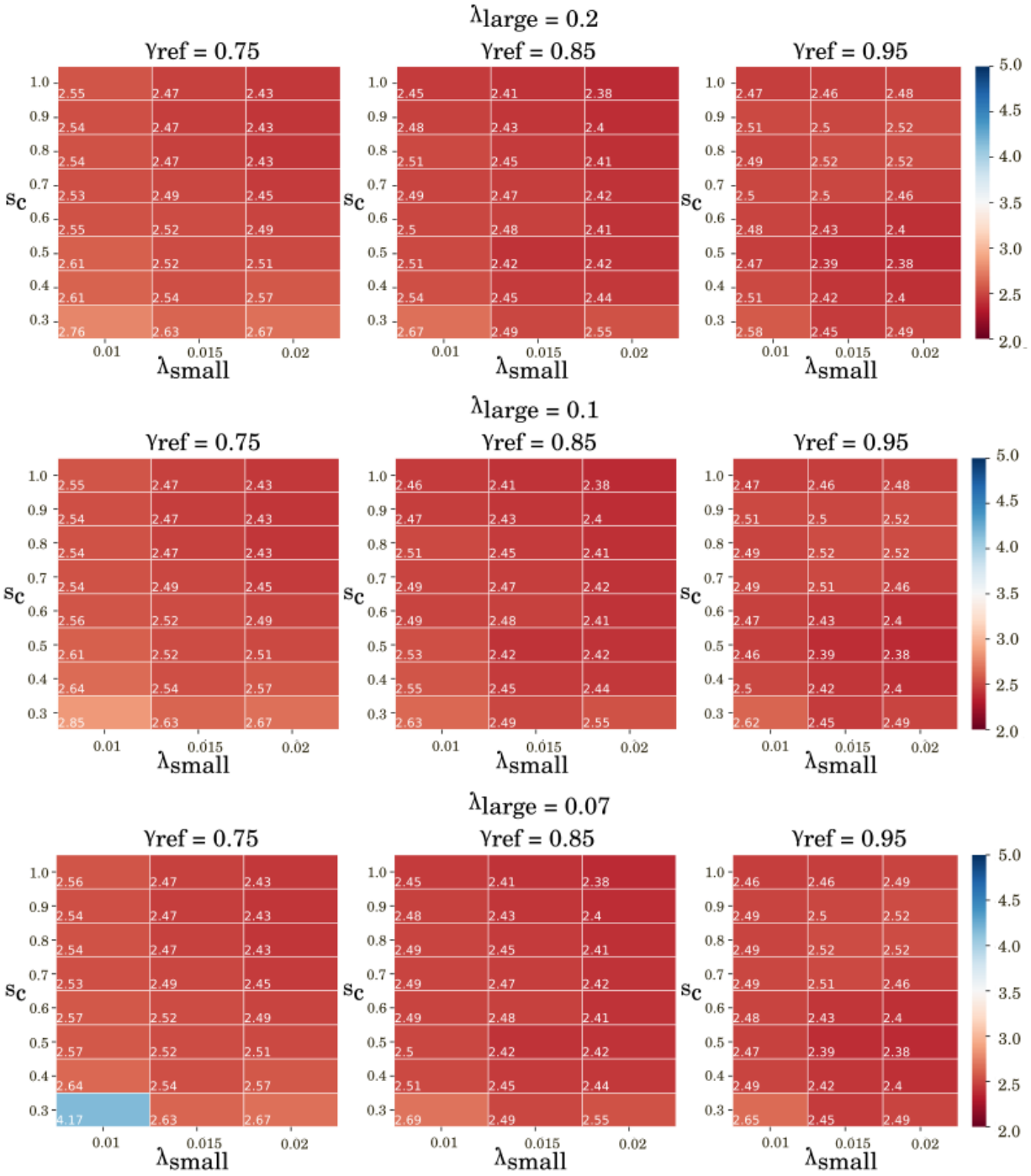


Figure 7: Heat map of mean absolute error (MAE) for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the D1200 dataset. All underlying NCI indices calculations were carried out using the promolecular approach and errors are in kJ/mol.

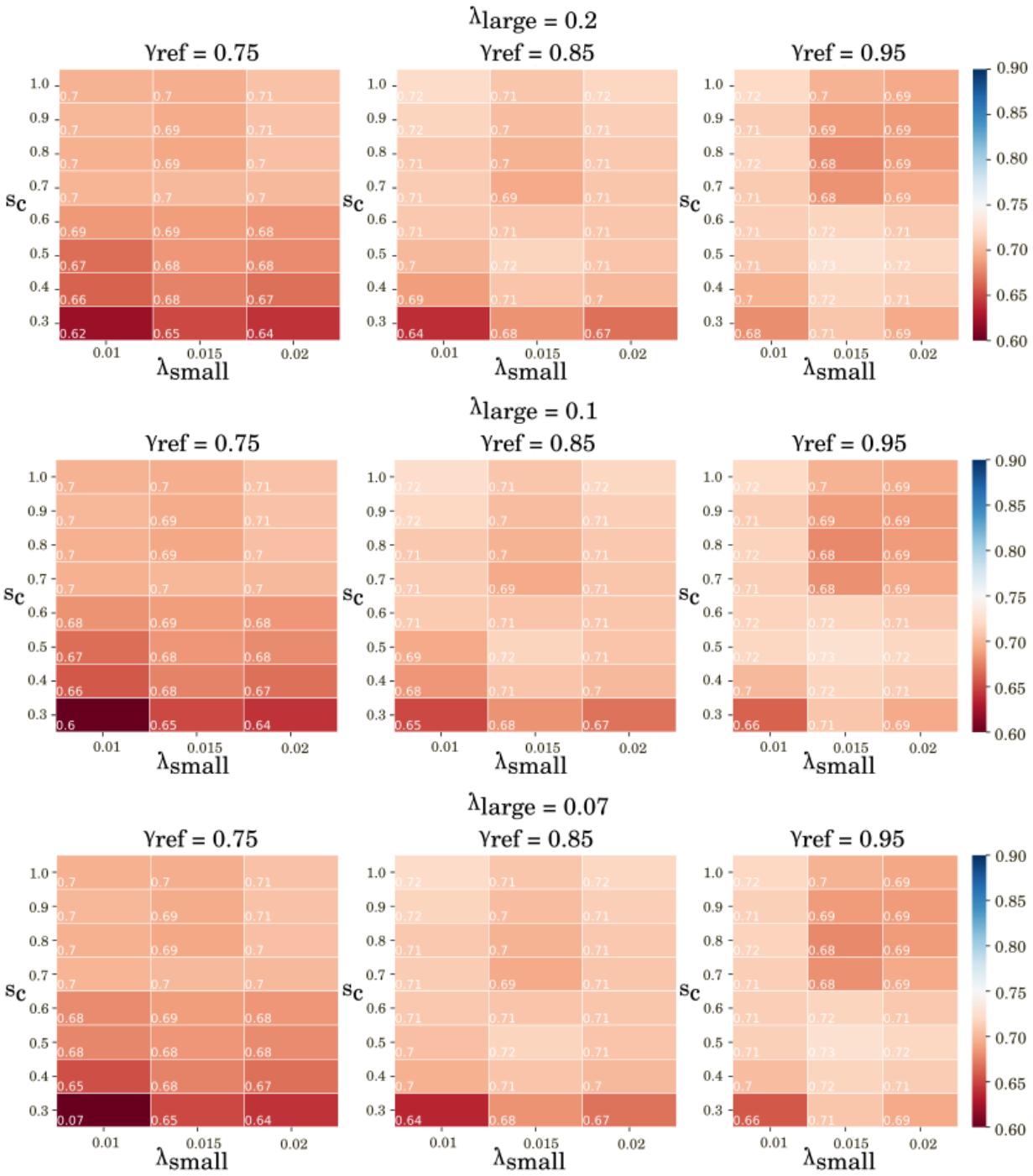


Figure 8: Heat map of R^2 for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the D1200 dataset. All underlying NCI indices calculations were carried out using the promolecular approach.

1.5 Parametrisation of HB375 dataset with DFT densities

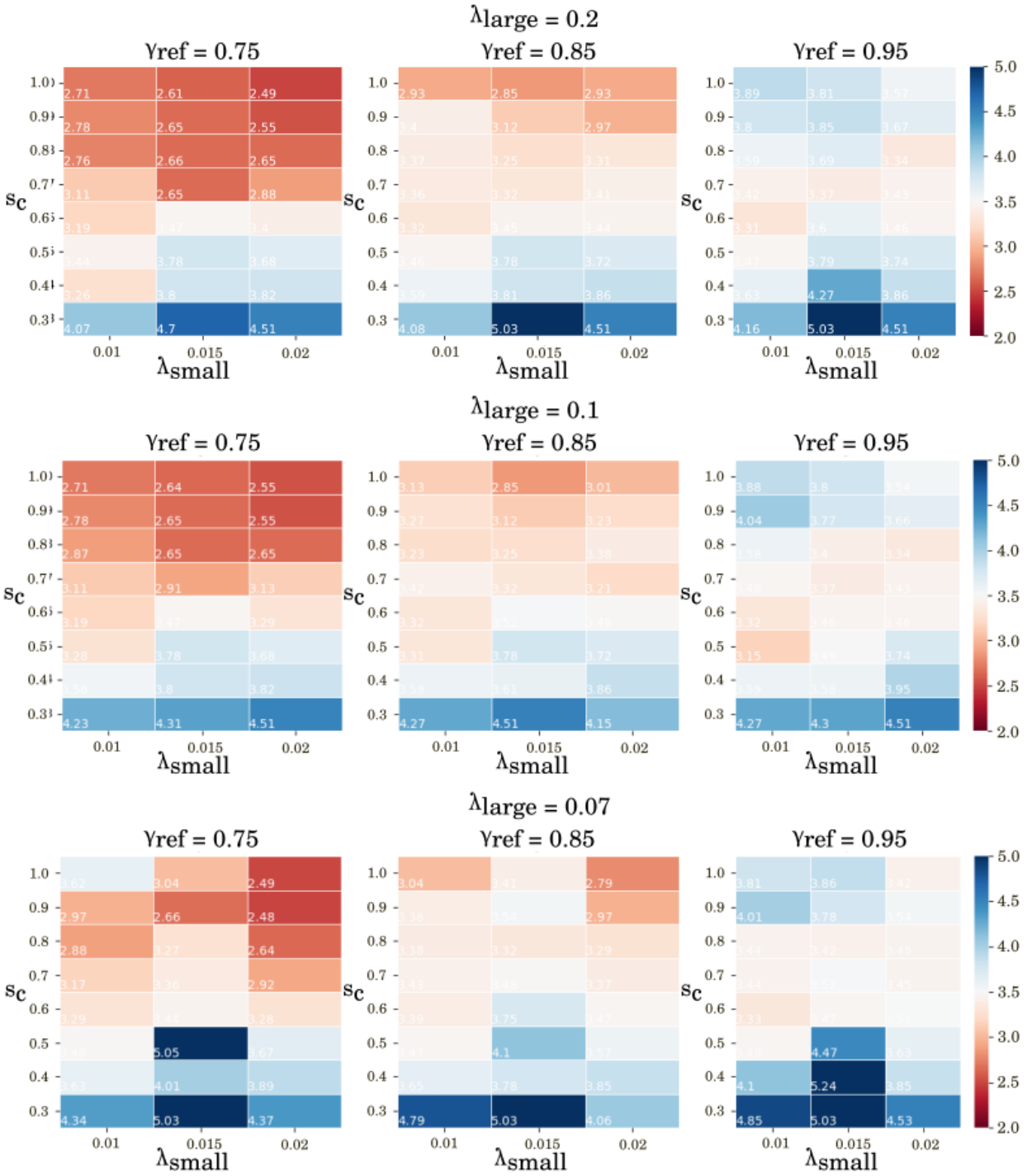


Figure 9: Heat map of mean absolute error (MAE) for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the HB375 dataset. All underlying NCI indices calculations were carried out using DFT densities and errors are in kJ/mol.

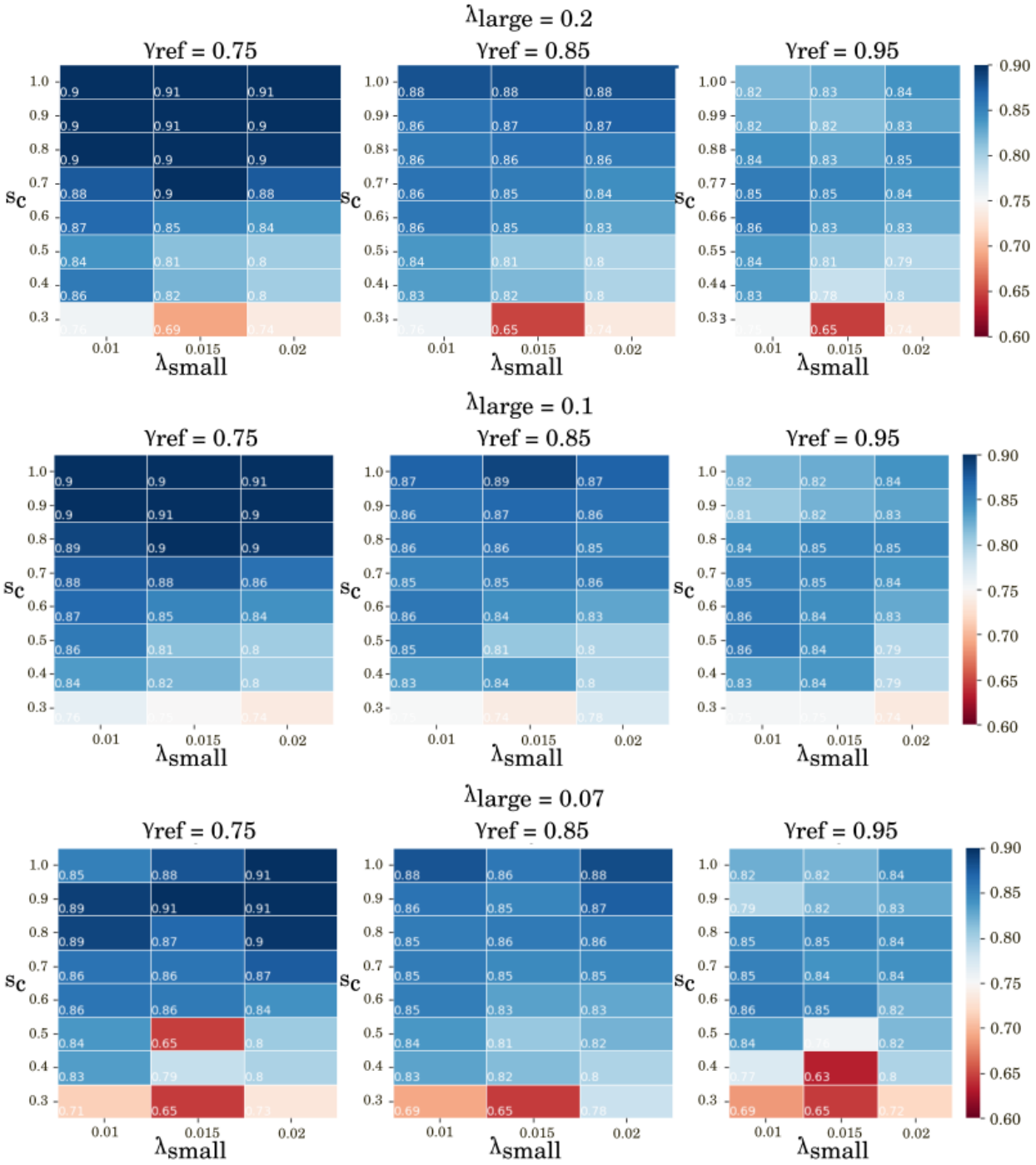


Figure 10: Heat map of R^2 for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the HB375 dataset. All underlying NCI indices calculations were carried out using DFT densities and errors are in kJ/mol.

1.6 Parametrisation of D1200 dataset with DFT densities

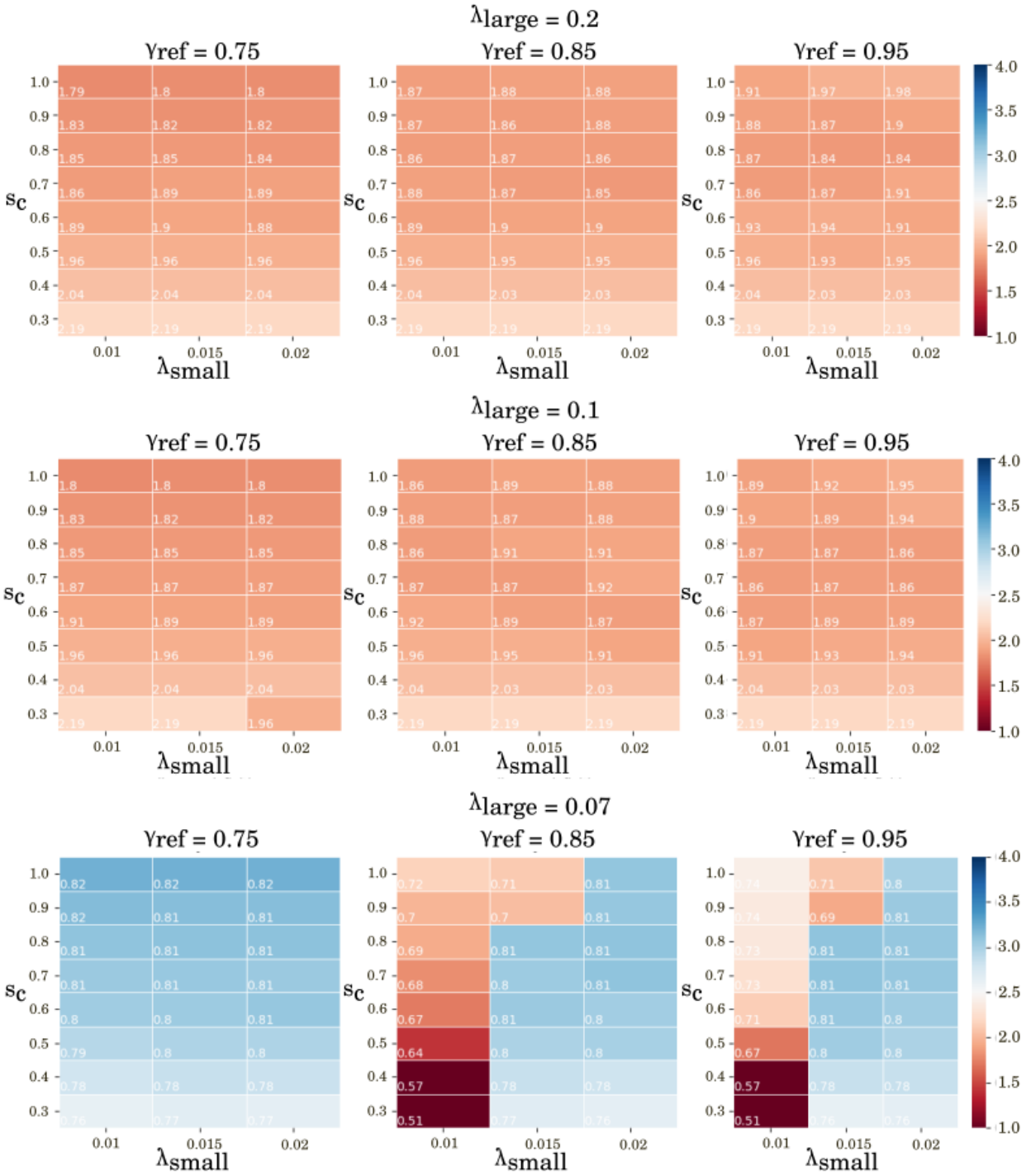


Figure 11: Heat map of mean absolute error (MAE) for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the D1200 dataset. All underlying NCI indices calculations were carried out using DFT densities and errors are in kJ/mol.

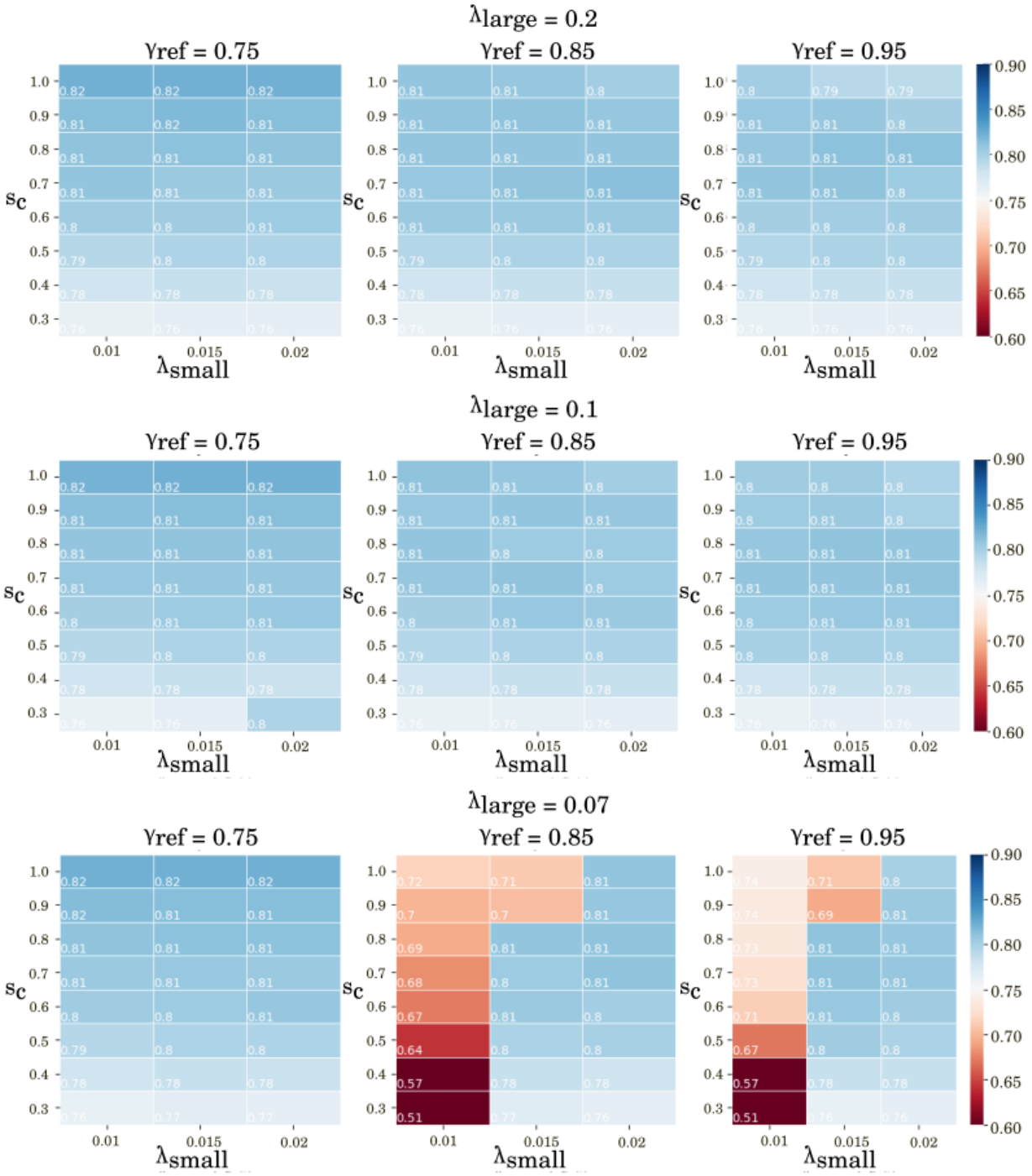


Figure 12: Heat map of R^2 for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the D1200 dataset. All underlying NCI indices calculations were carried out using DFT densities and errors are in kJ/mol.

2 Multivariate Gradient Boosting Regression

2.1 Internal regression parameters

The following are sklearn gradient boosting regression parameters which were quickly optimized to prevent under- and over-fitting, and excessively long calculations. Train-test split: test_size=0.33, random_state=42) Regression parameters: random_state=None, learning_rate=0.1 ,max_depth=3, n_estimators=100, max_features=None)

2.2 Parametrisation of HB375 dataset with promolecular densities

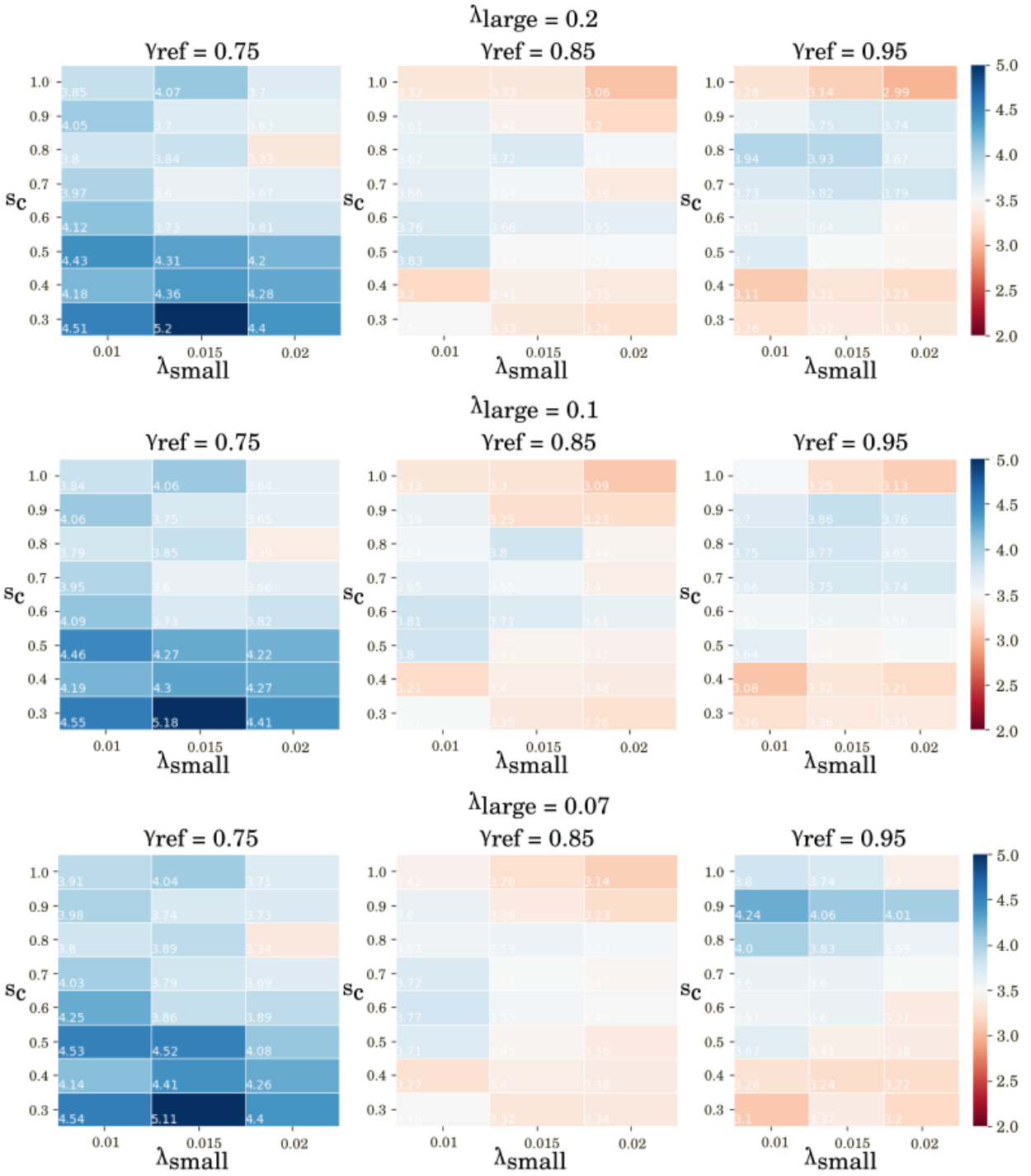


Figure 13: Heat map of mean absolute error (MAE) for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.05$ for the HB375 dataset. All underlying NCI indices calculations were carried out using the promolecular approach and gradient boosting regression and errors are in kJ/mol.

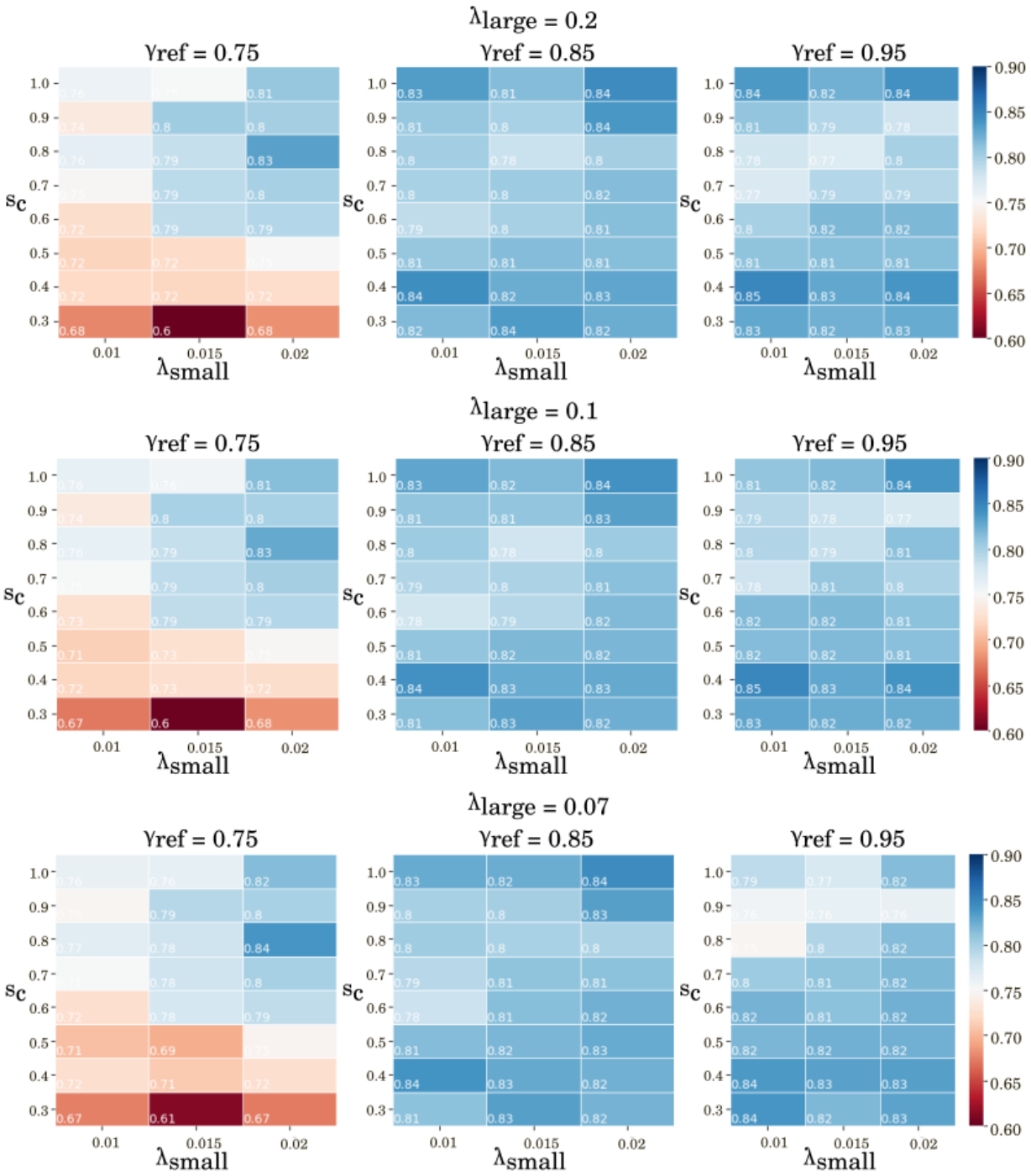


Figure 14: Heat map of R^2 for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.05$ for the HB375 dataset. All underlying NCI indices calculations were carried out using the promolecular approach and gradient boosting regression.

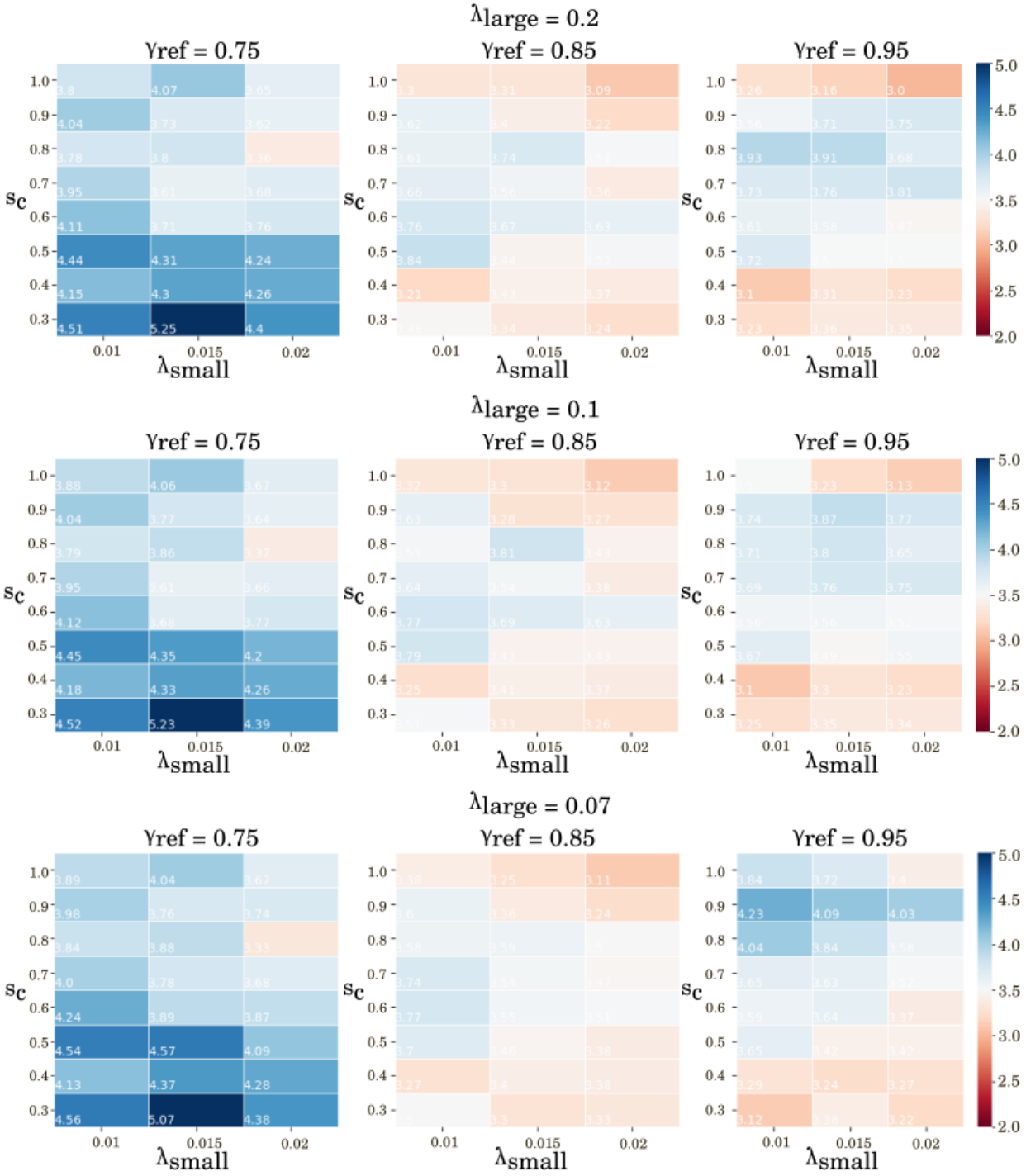


Figure 15: Heat map of mean absolute error (MAE) for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the HB375 dataset. All underlying NCI indices calculations were carried out using the promolecular approach and gradient boosting regression and errors are in kJ/mol.

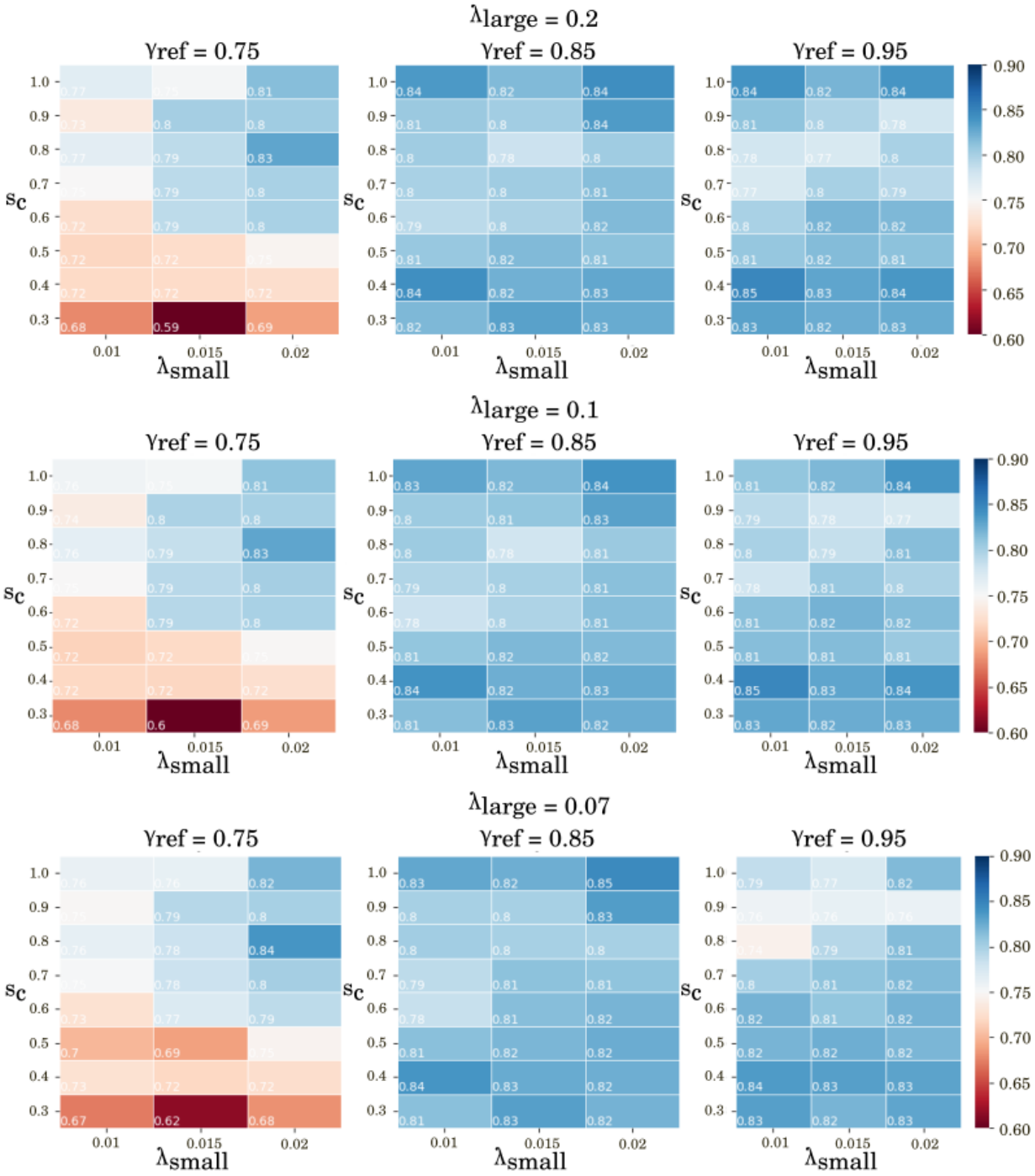


Figure 16: Heat map of R^2 for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the HB375 dataset. All underlying NCI indices calculations were carried out using the promolecular approach and gradient boosting regression.

2.3 Parametrisation of D1200 dataset with promolecular densities

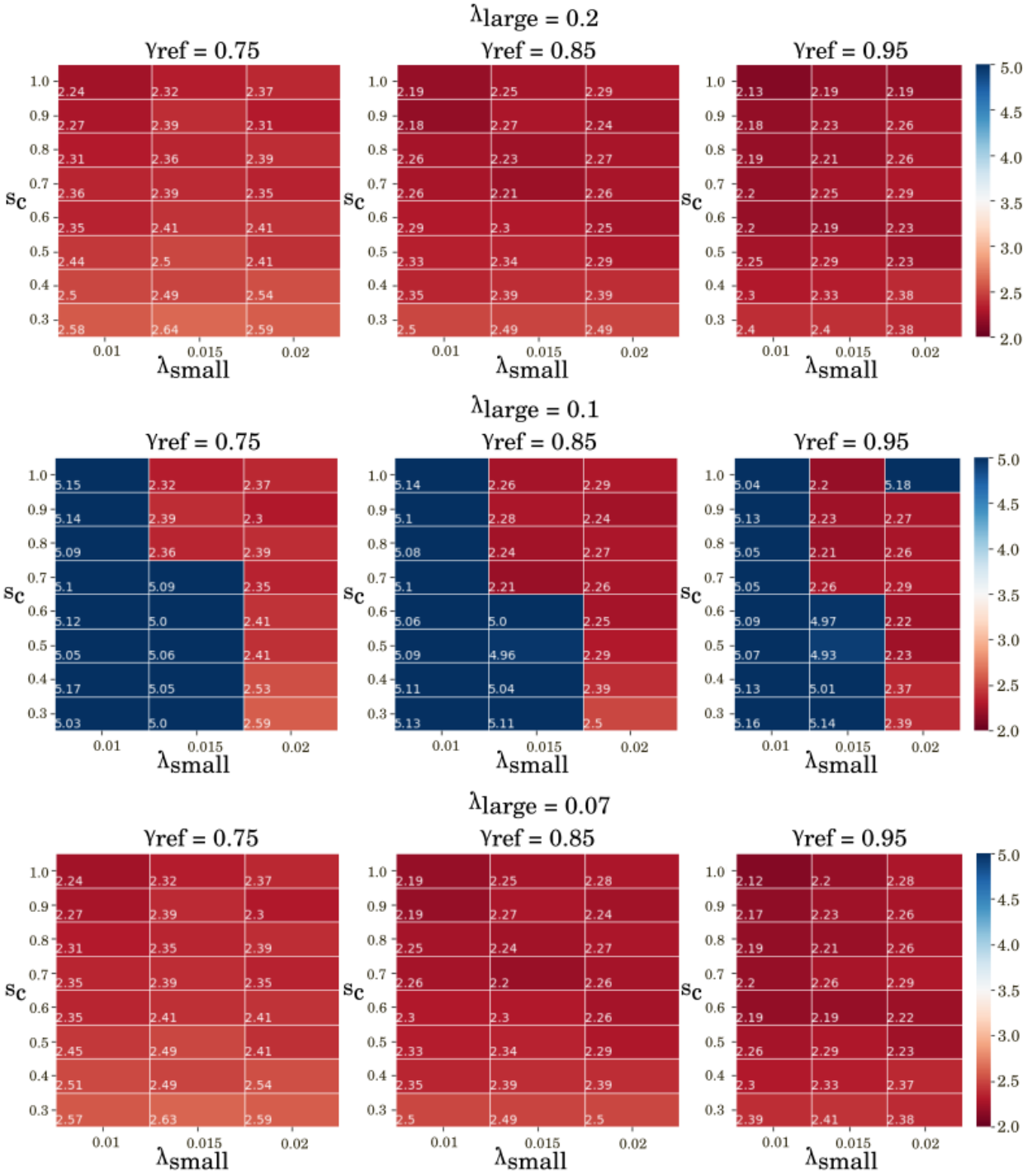


Figure 17: Heat map of mean absolute error (MAE) for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.05$ for the D1200 dataset. All underlying NCI indices calculations were carried out using the promolecular approach and gradient boosting regression and errors are in kJ/mol.

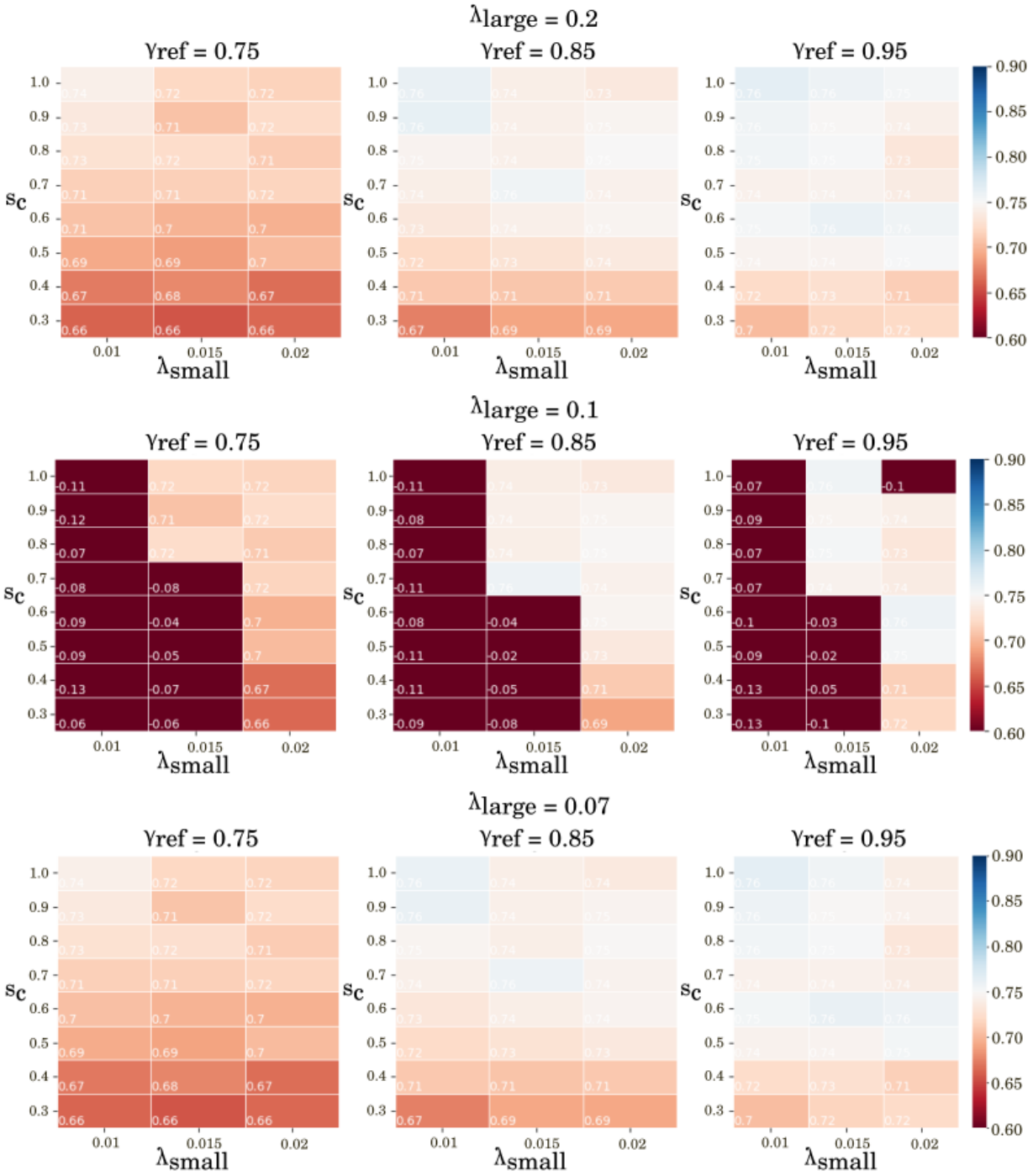


Figure 18: Heat map of R^2 for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.05$ for the D1200 dataset. All underlying NCI indices calculations were carried out using the promolecular approach and gradient boosting regression.

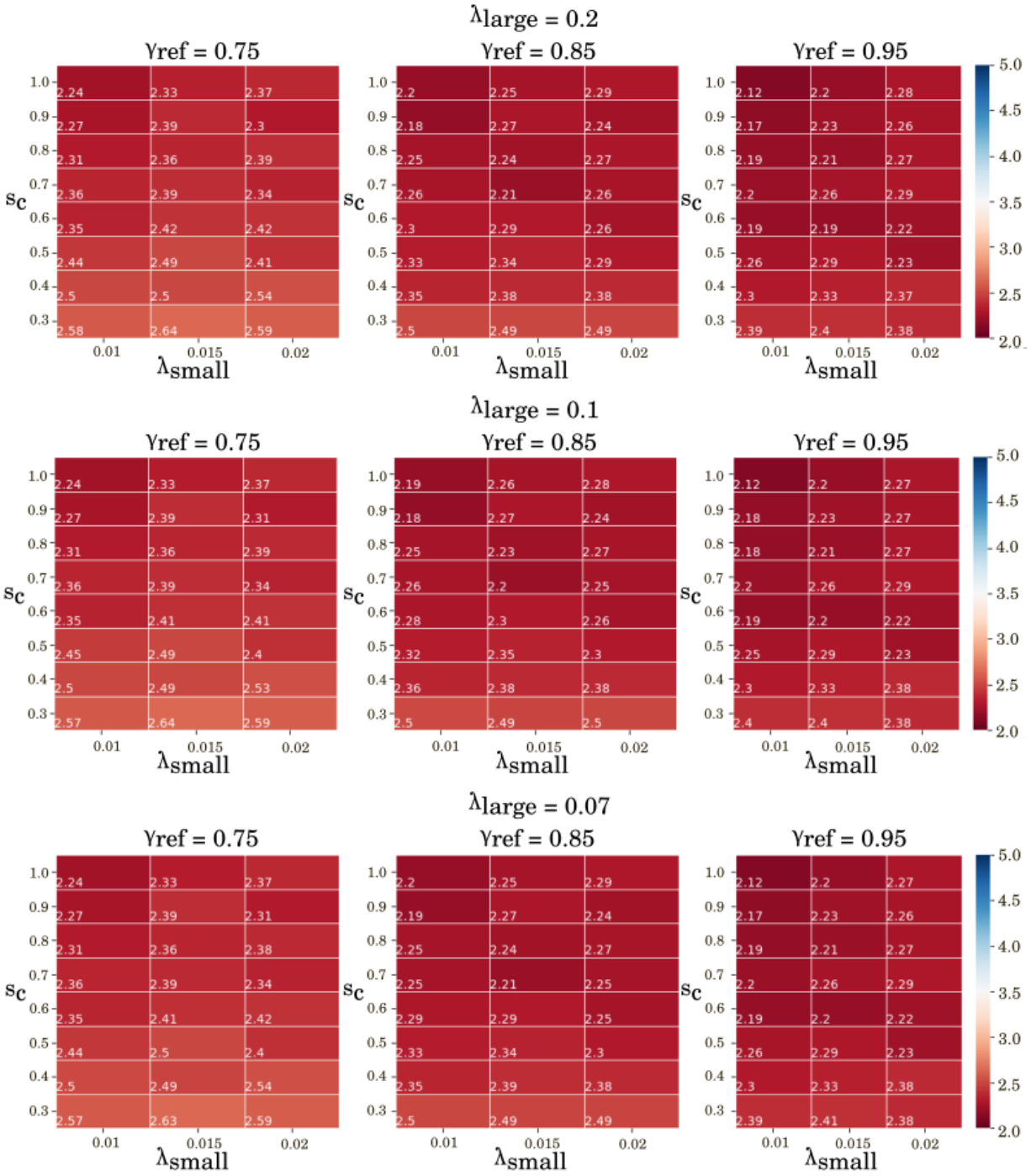


Figure 19: Heat map of mean absolute error (MAE) for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the HB375 dataset. All underlying NCI indices calculations were carried out using the promolecular approach and gradient boosting regression and errors are in kJ/mol.

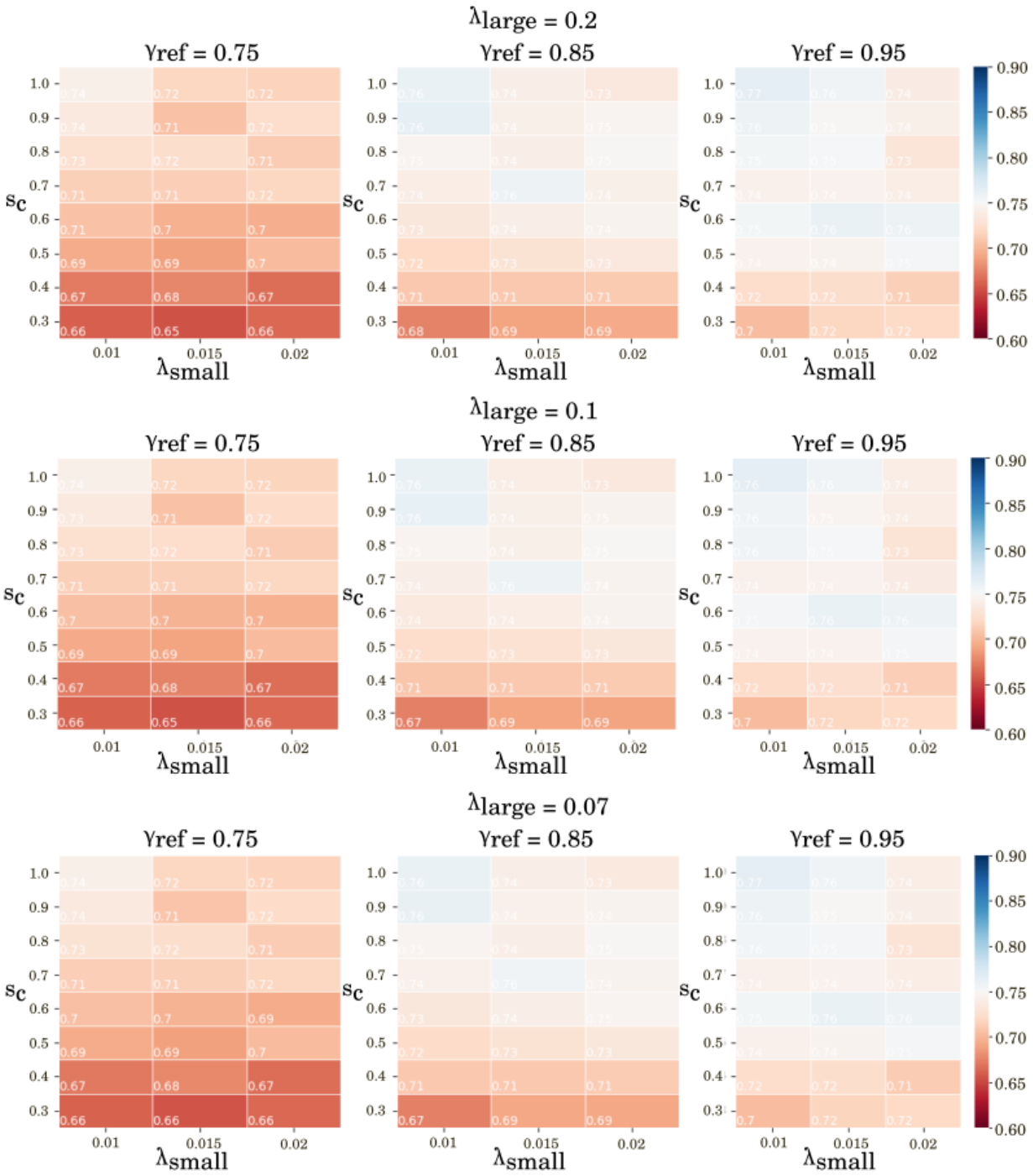


Figure 20: Heat map of R^2 for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the HB375 dataset. All underlying NCI indices calculations were carried out using the promolecular approach and gradient boosting regression.

2.4 Parametrisation of HB375 dataset with DFT densities

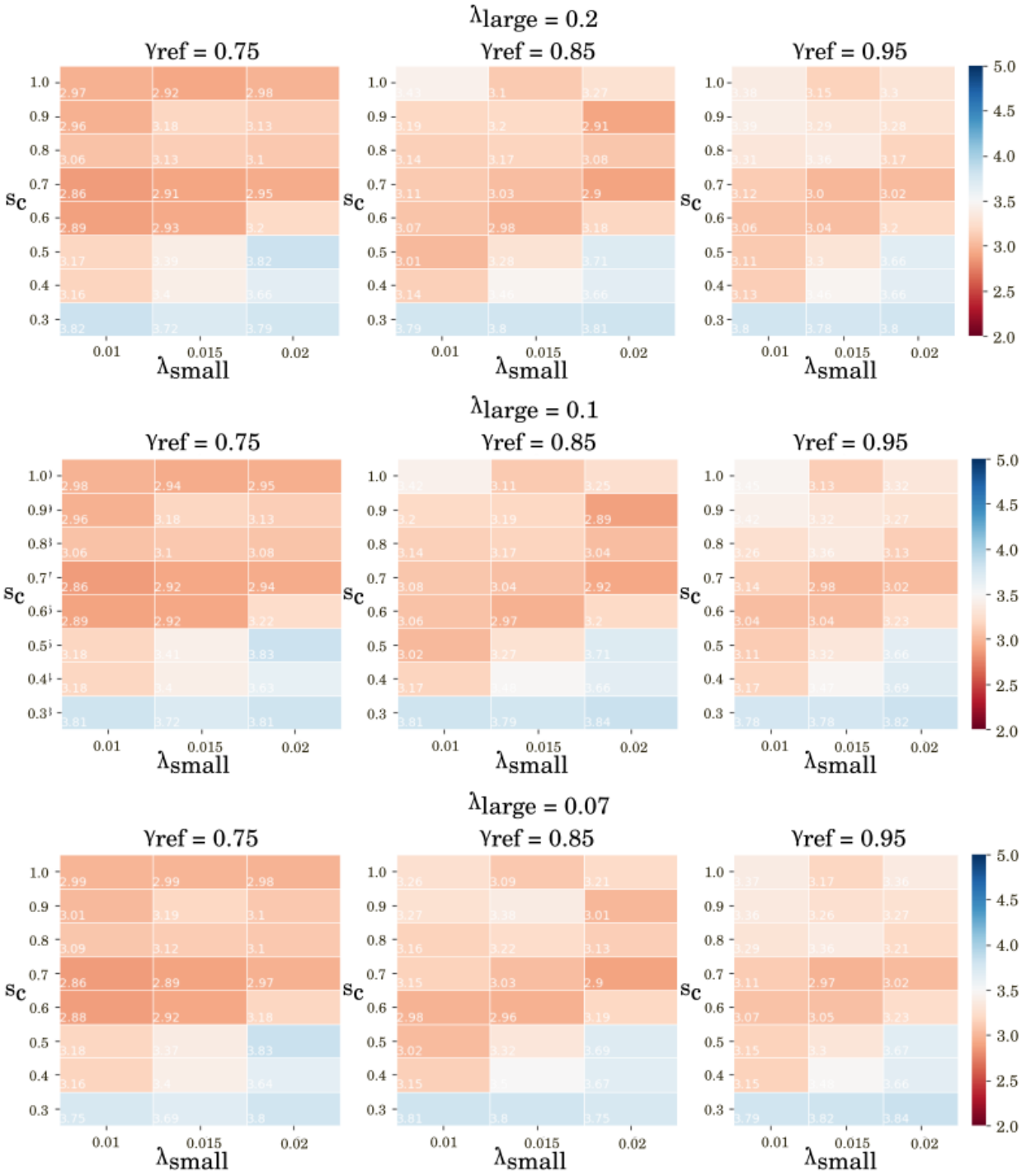


Figure 21: Heat map of mean absolute error (MAE) for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the HB375 dataset. All underlying NCI indices calculations were carried out using DFT densities and gradient boosting regression and errors are in kJ/mol.

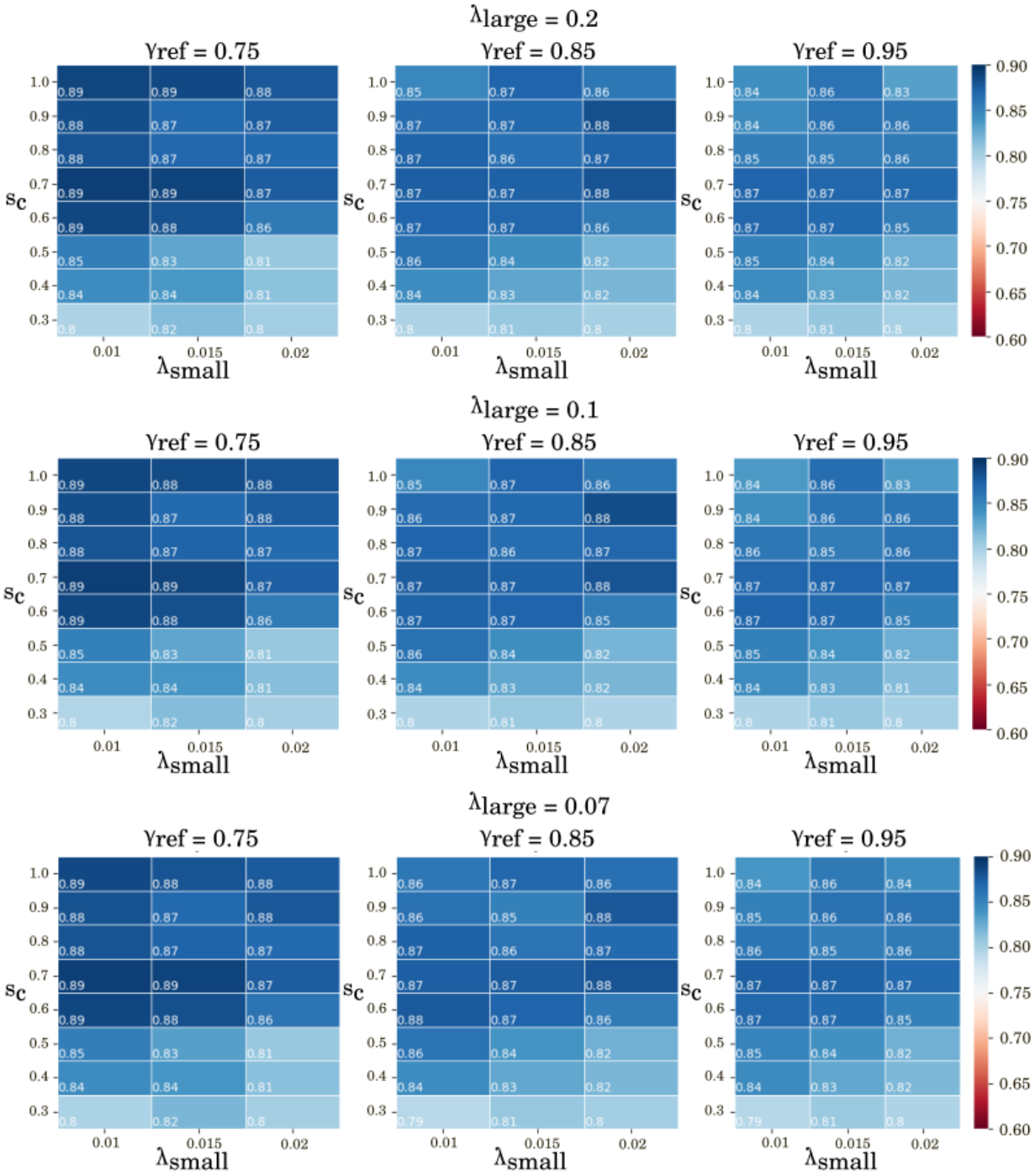


Figure 22: Heat map of R^2 for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the HB375 dataset. All underlying NCI indices calculations were carried out using DFT densities and gradient boosting regression.

2.5 Parametrisation of D1200 dataset with DFT densities

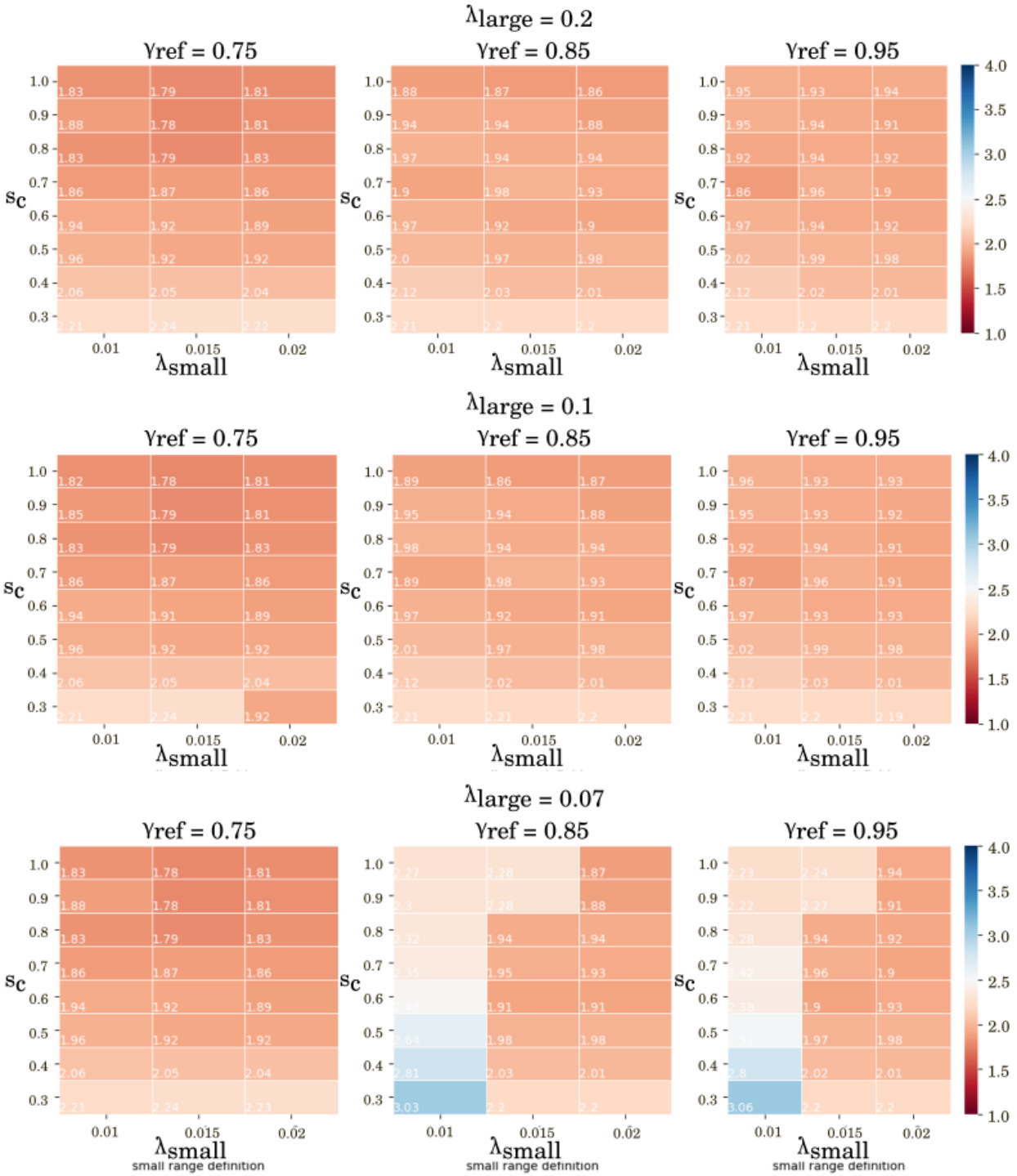


Figure 23: Heat map of mean absolute error (MAE) for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the D1200 dataset. All underlying NCI indices calculations were carried out using DFT densities and gradient boosting regression and errors are in kJ/mol.

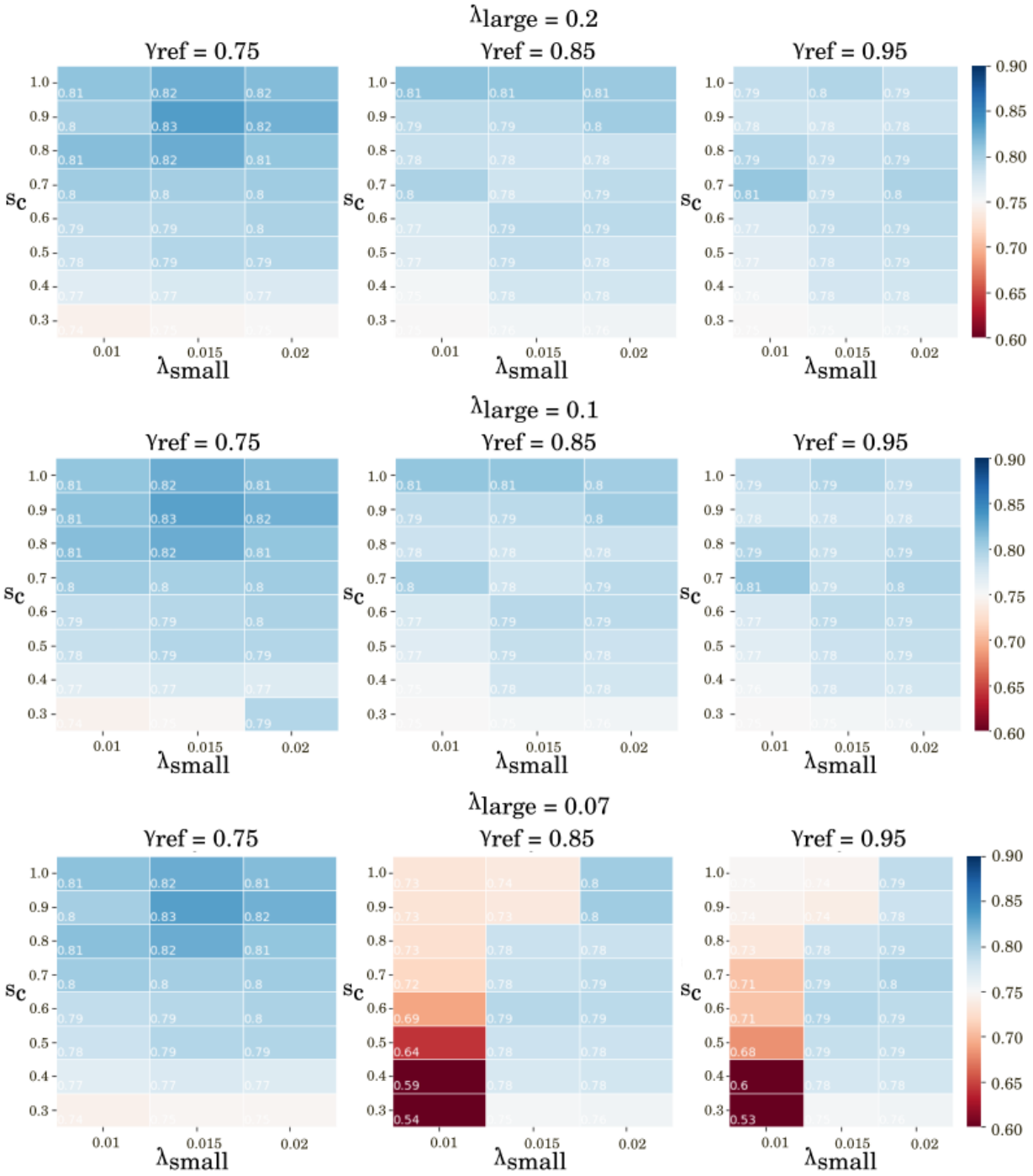


Figure 24: Heat map of R^2 for two varying parameters: s_c and λ_{small} for specified λ_{large}, and γ_{ref} and $\rho_c = 0.07$ for the D1200 dataset. All underlying NCI indices calculations were carried out using DFT densities and gradient boosting regression.

3 HB375 Internal not-hydrogen-bonded control results

The HB375 dataset contained a set of 108 "no hydrogen bond" compounds - ones which did not feature a conventional hydrogen bond (a short contact between N/O and H bound to N/O) but were still found to be primarily driven by electrostatics as found by SAPT analysis[?]

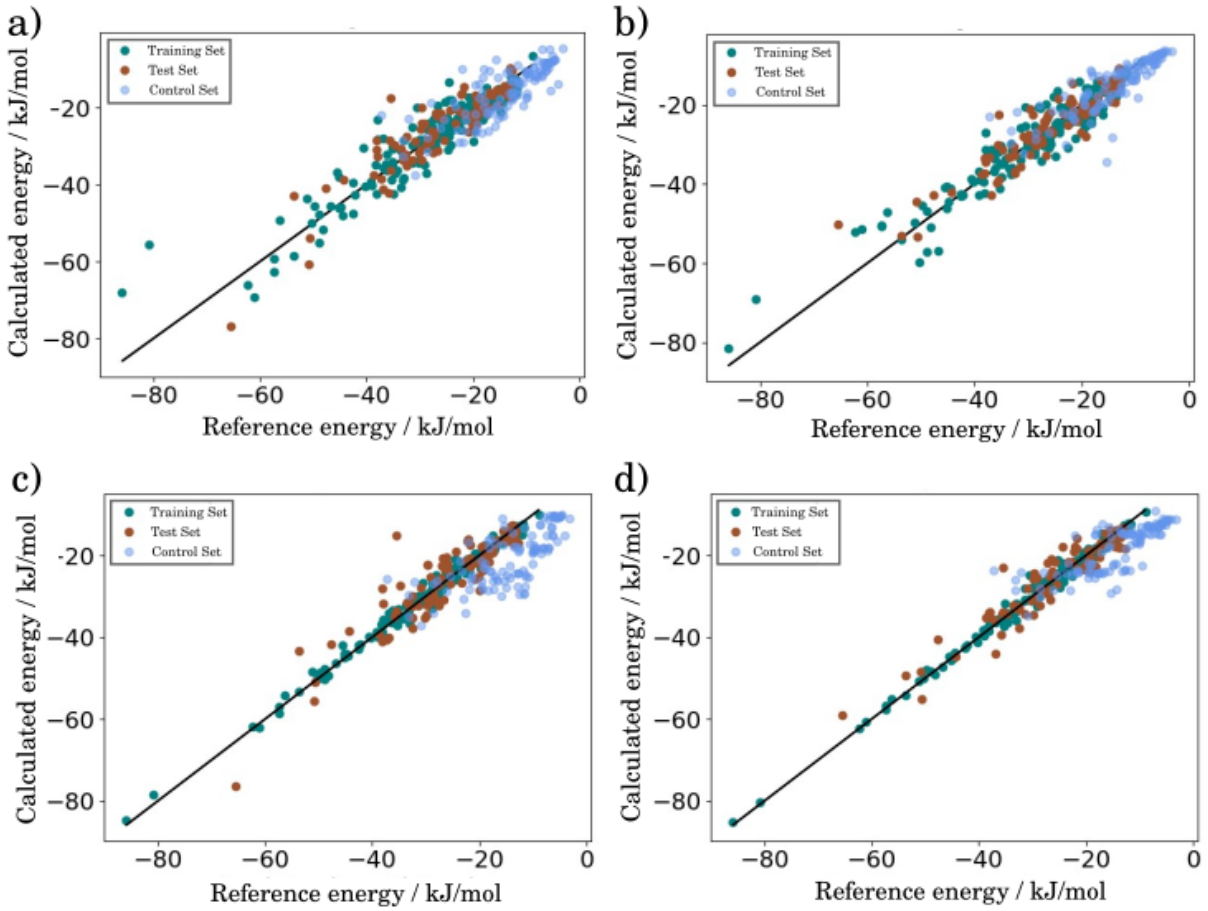


Figure 25: a) promolecular symbolic regression, b) promolecular gradient boosting regression, c) DFT symbolic regression, d) DFT gradient boosting regression calculated for the HB375 dataset with optimized parameters. For the regression calculation, the dataset was split into train and test sets at a 2:1 ratio and the 108 control compounds were separated completely and plotted with the obtained model.

4 Optimized results using the DFT densities

An analysis parallel to the one presented in the main article text was performed here but for the NCIPLOT calculations done with the DFT densities. The DFT HB375 equation using the best parameters - ($\lambda_{large} = 0.2$, $\lambda_{small} = 0.02$, $\rho_c = 0.07$, $s_c = 1.0$, and $\gamma_{ref} = 0.75$ - is

$$E_{Hydrogen\_bond}(\rho) = -(3.4 \times 10^3 \left(I_{3/2,van\_der\_Waals} + I_{3/2,Hydrogen\_bond}\right) + 1.15 \times 10^2 \sqrt[3]{I_{5/2,van\_der\_Waals}}) \tag{7}$$

and gave $R^2 = 0.91$ and MAE = 2.51 kJ/mol for the test set.

Similarly, DFT D1200 equation using the best parameters - $\lambda_{large} = 0.2$, $\lambda_{small} = 0.02$, $\rho_c = 0.07$, $s_c = 1.0$, and $\gamma_{ref} = 0.75$ - is

$$E_{van\_der\_Waals}(\rho) = -\left(8.1 \times 10^2 \times I_{1,van\_der\_Waals}\left(-I_{1,van\_der\_Waals} + 0.2966816\right) + 1.3436404\right) \tag{8}$$

and gave $R^2 = 0.825$ and MAE = 1.80 kJ/mol for the test set.

The best resultant equation for the DFT-density-derived NCI approach was not a direct sum of equations (1) and (2) but instead, appears to be a more complicated combination, whereby only the repulsive component from equation (2) was used, suggesting a subtle effect that was not captured by equation (1). For the promolecular approach, the joined equation used an entirety of the D1200 equation for the van der Waals equations, however, such a combined equation (and indeed all other combinations) produced many worse-performing equations. Hence, the final equation was

$$E_{int}(\rho) = -(3.4 \times 10^3 \left(I_{3/2,van\_der\_Waals} + I_{3/2,Hydrogen\_bond}\right) + 1.15 \times 10^2 \sqrt[3]{I_{5/2,van\_der\_Waals}} - 8.1 \times 10^2 \times I^2_{1,van\_der\_Waals}) \tag{9}$$

which gave a small improvement over the promolecular result. The performance of this

combined equation (3) is shown in Figure 26 b) and gave an improved R^2 of 0.89 and MAE of 2.20 kJ/mol.

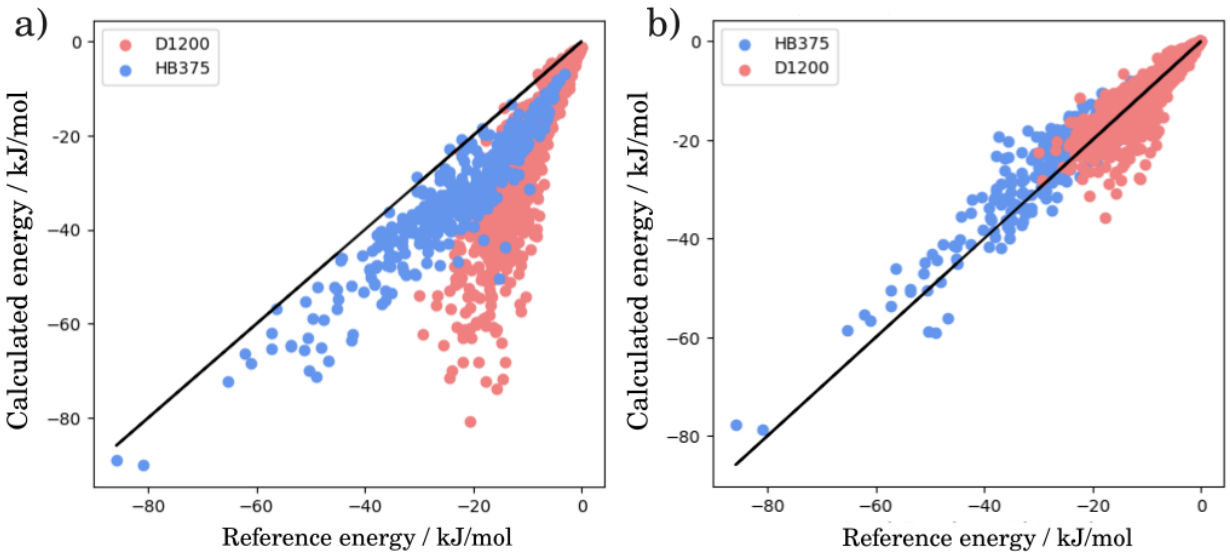


Figure 26: Scatter plot of calculated energy using a single equation from DFT-derived densities versus CCSD reference for the color-coded datasets: blue is HB375 and pink is D1200. a) represents the exact sum of equations (1) and (2), and b) represents equation (3). The overall R^2 and MAEs are: a) 0.66 and 12.17 kJ/mol, and b) 0.89 and 2.20 kJ/mol.

As the equation changed with the use of DFT versus the promolecular approach, it also was sensitive to the accuracy of electron density used. We wondered if this would be functional-dependent, and hence performed a calculation with B3LYP-D3, PBE0-D3, and SVWN functionals (and def2-SVP basis set, and the same NCIPLOT parameters for all) to evaluate the possible differences. As seen in Figure 27, the differences due to the chosen functional are insignificant.

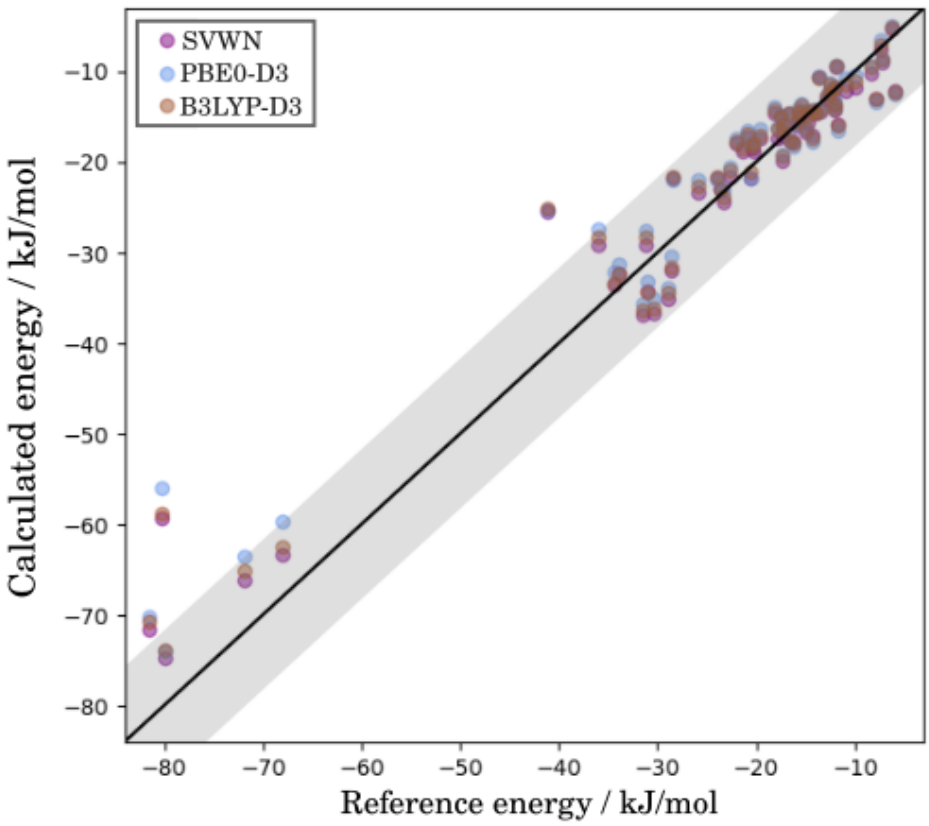


Figure 27: Scatter plot of calculated energy using a single equation from DFT-derived densities versus CCSD reference for the S66 dataset: color-coded by functional used in the calculation. The R^2 and MAEs are: a) 0.93 and 2.86 kJ/mol, and b) 0.92 and 3.18 kJ/mol, c) 0.93 and 2.97 kJ/mol.

5 HB375x10 for extended and compressed geometries

The HB375x10 dataset contained a series of equilibrium as well as compressed and extended structures by a constant fraction for all the complexes. We wanted to check the extent to which the symbolic equation (7) derived in the main article text could be used to extrapolate to other geometries.

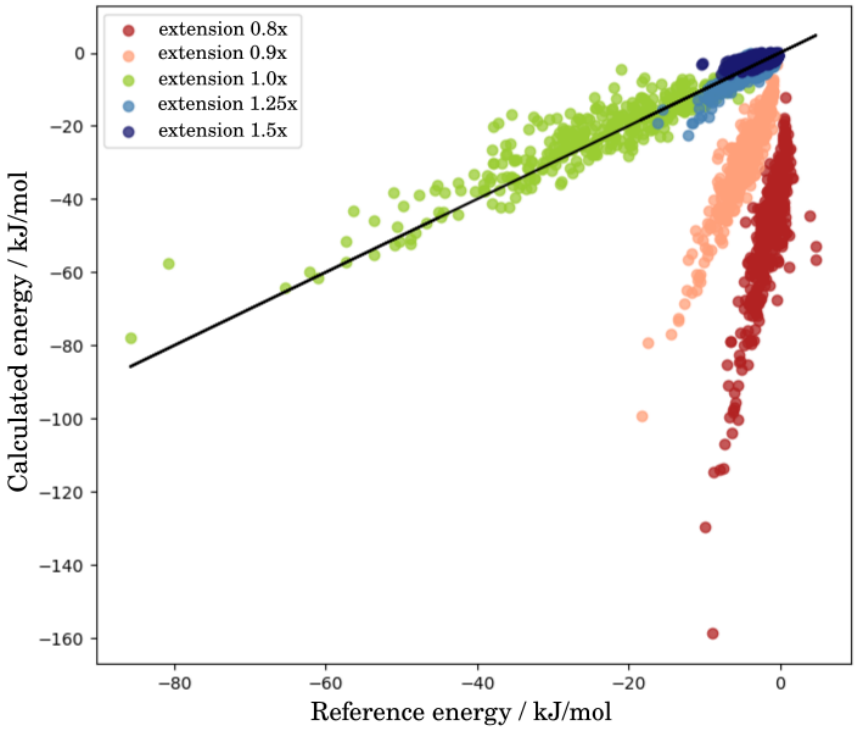


Figure 28: HB375 predicted binding energies calculated with the optimum equation (7) for 5 color-coded geometry extensions (0.8x, 0.9x, 1.0x, 1.25x, 1.5x). The prediction of equilibrium geometry (1.0x) is very good, yet predictions of out-of-equilibrium complexes, especially compressed geometries (¡1.0x) show straight lines with much steeper gradients.

Figure 28 shows the energies calculated with the same equation, and clearly showcases the limitation of the equation's lack of repulsion term. For small extensions, the energy predictions were fair, yet for larger extensions, the best-fit line would have a smaller gradient than the y=x line suggesting the NCI indices were dropping faster than expected with the increasing distance. For compressions, the decreasing distance increased the NCI indices for hydrogen bonding and van der Waals, but with no repulsive compensation, the energies increased dramatically. Interestingly, the energy predictions could still be fit to a straight line, but with a differing gradient. Such a line could be found by retraining the symbolic regression model at different extensions, but clearly, such an approach does not generalize easily. The other route would be to include the ratio of extension in the training information and, therefore, use the entirety of the data to identify one equation. This does produce an equation including the extension variable; however, such an equation is incapable of

predicting the energy well for an array of extensions, see Figure 29.

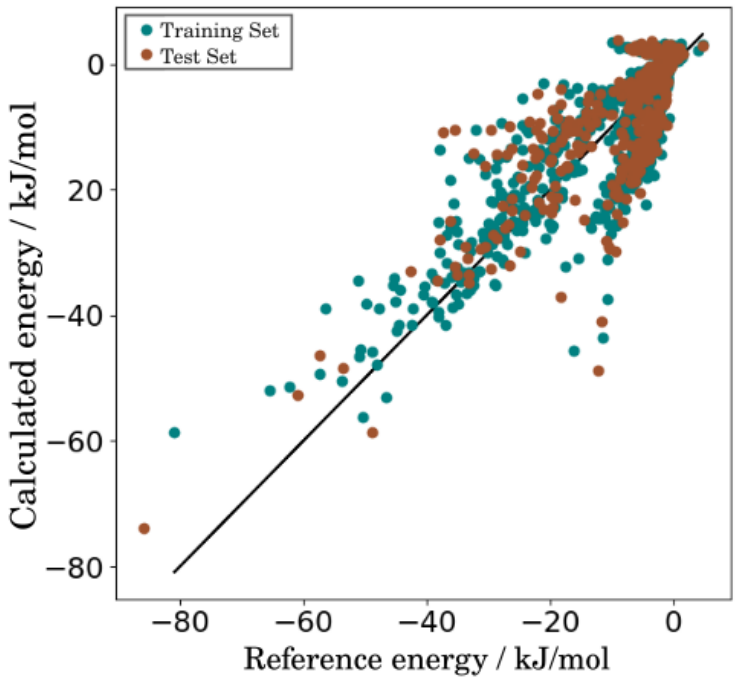
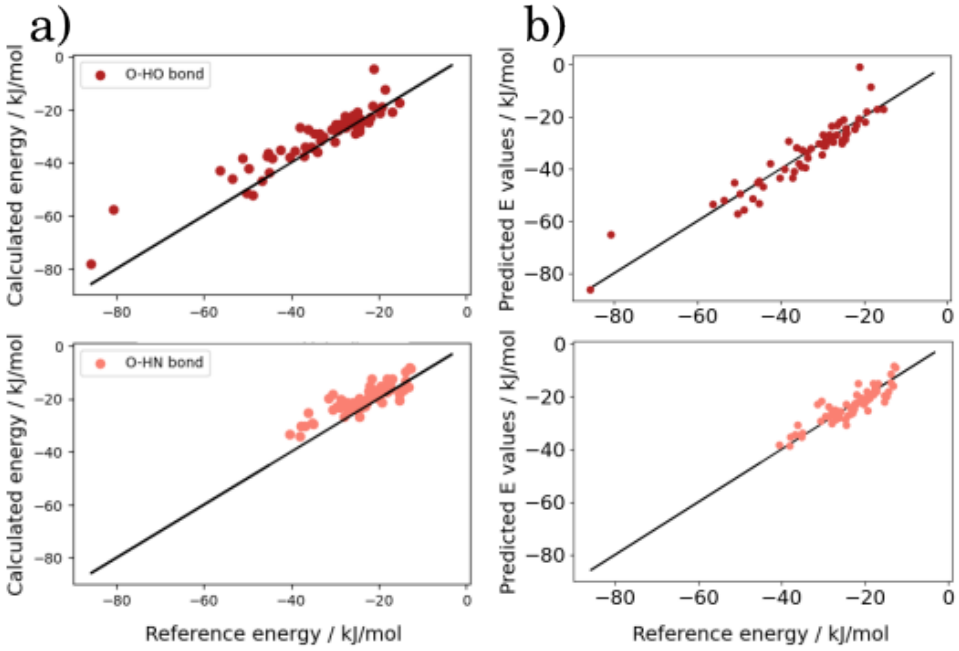


Figure 29: HB375 dataset at 5 geometry extensions (0.8x, 0.9x, 1.0x, 1.25x, 1.5x) modeled with symbolic regression with PySR. The input data included the extension, and NCI indices of hydrogen bond, van der Waals and repulsion contacts. The train:test split was 2:1.

6 Regression for subsets of the HB375 and D1200

Both HB375 and D1200 datasets are meant to represent the chemical space and, hence, contain a variety of interaction subtypes; specifically, HB375 can be divided by bond type - depending on the identity of the electron donor (OH, NH, CH) and acceptor (O, C, π system) - and D1200 can be divided based on elements present in the monomers - second and third row, halides, and noble gases. As such, the underlying interactions could differ in the balance of underlying energetic contributions and be best represented by a slightly different symbolic equation. Therefore, to test the extent of this effect, we created subsets of each of the datasets, fitted them with PySR separately and compared the results. HB375 results are present in Figure 30 and D1200 results in Figure 31. Qualitatively, the fits do not look much improved, showing much the same scatter. Quantitatively, the overall HB375 fit gave R^2 =

0.86 and MAE = 3.35 kJ/mol, versus the separated fits which gave improved results of R^2 = 0.90 and MAE = 2.87 kJ/mol for the re-combined dataset. For D1200, the original equation gave R^2 = 0.75 and MAE = 2.22 kJ.mol, whereas the re-combined dataset gave R^2 = 0.785 and MAE 2.01 = kJ/mol. The overall errors and correlations might have been improved indeed, with such a slight difference, that we feel confident concluding the equations (5) (6) and (7) in the main text represent the variability of the hydrogen bonds and dispersion interactions sufficiently. Furthermore, the implementation of unique parametrisation for each interaction subtype would vastly complicate the underlying NCIPLOT algorithm and likely contribute to worsening the computation time.



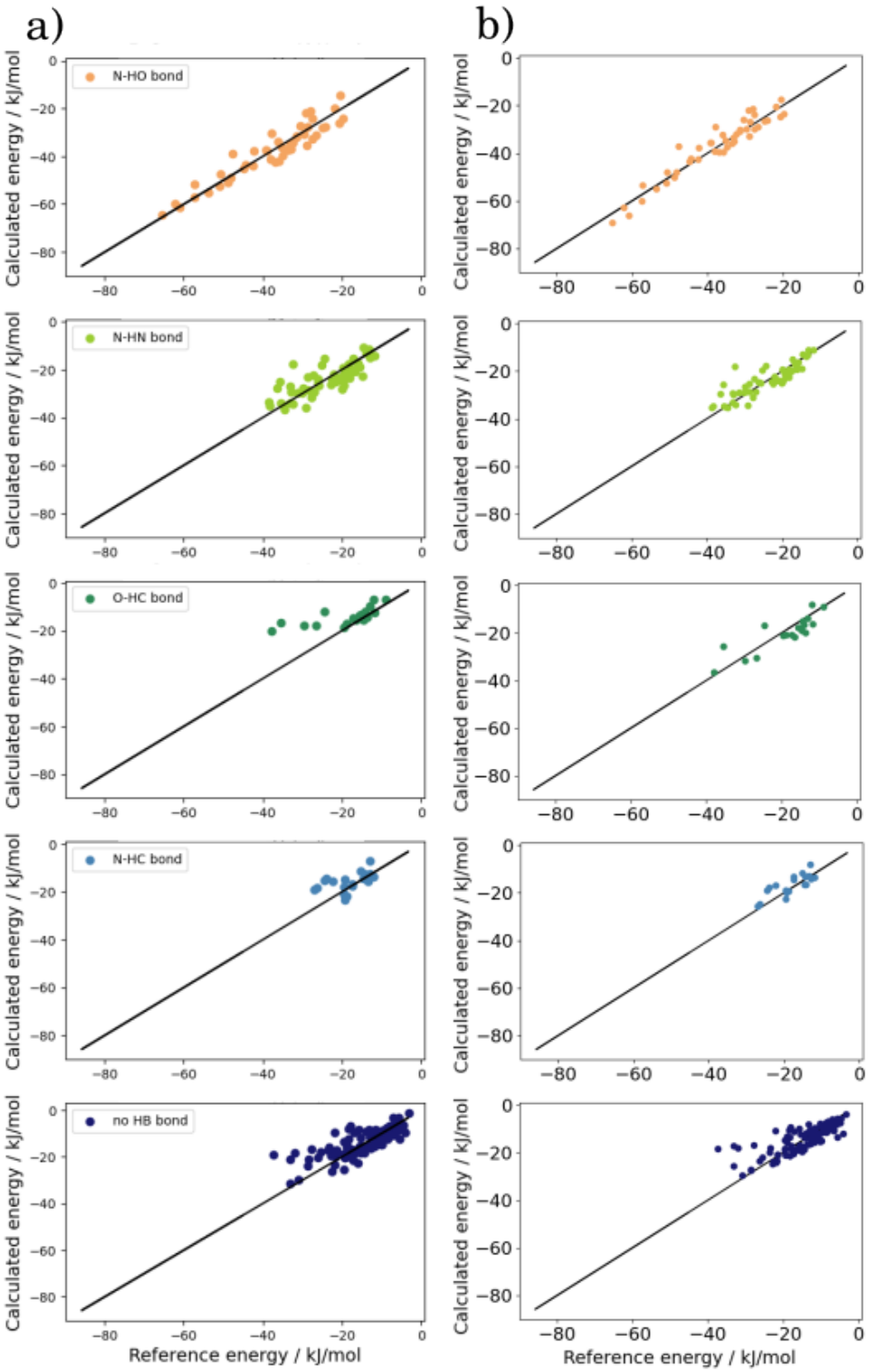


Figure 30: HB375 dataset separated by the nature of hydrogen bond donor and acceptor. a) Predicted binding energy using the equation (7) in the main text, b) Separately fitted equation for the subset. All results used the promolecular approach.

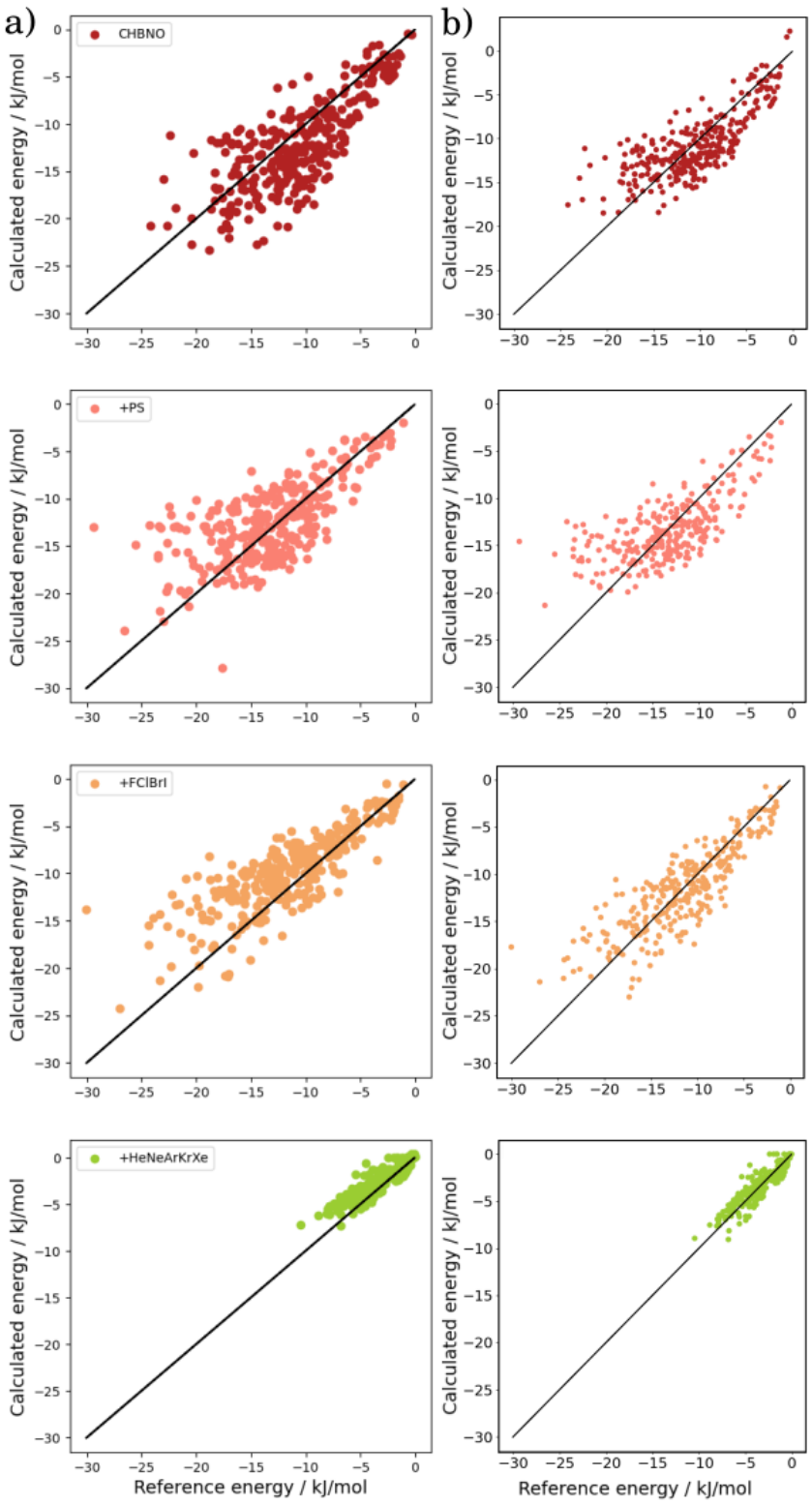


Figure 31: D1200 dataset separated by the nature of hydrogen bond donor and acceptor. a) Predicted binding energy using the equation (7) in the main text, b) Separately fitted equation for the subset. All results used the promolecular approach.