Quantitative NCI index: Defining the link between NCI index and interaction energy

Katarzyna Kate Zator, Julia Contreras-García

Quantitative NCI index: Defining the link between NCI index and interaction energy

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

†*Laboratoire de Chimie Théorique, Sorbonne Université and CNRS, 4 Pl Jussieu, 75005, Paris, France*

‡*CNRS*

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

Phone: +33 1 44 27 38 79

Abstract

The NCI method has existed for over a decade, and it has had great success in qualitative studies of non-covalent interactions in a variety of systems. It relies on the detection of low-density and low-reduced density gradient volumes in intermolecular complexes that co-locate at sites of interactions making for an excellent visualization tool.

A quantitative aspect has been theorized as the integral over the detected volume, though the derivation of the equation justifying is still elusive. Consistent correlations between NCI indices and binding energies nonetheless suggest its existence. This work sets out to find the symbolic form of this relationship and evaluate its accuracy in the prediction of interaction energies in small dimeric complexes. To that end, it systematically evaluates the integration of NCI volumes, and their dependence on variable parameters, and hence clarifies the definitions of low densities. The symbolic relationship between the NCI index values and gold standard binding energies has been determined for the NCI atlas' hydrogen bond and dispersion datasets and calculated

the errors below 3.0 kJ/mol. The resultant equation was laterally tested on the S66 dataset and found to calculate its interaction energies at chemical accuracy. This work is the first step for the creation of a binding energy predictor from electron density alone across different chemical families, therefore for the creation of a tool capable of predicting and calculating NCIs in a variety of structures.

1 Introduction

Non-covalent interactions (NCIs) are the underlying reason behind an array of physical and biological processes and dictate the structure and function of biological macromolecules, their interactions with their environment, pharmaceuticals, agrochemicals, and materials.[1–5] Furthermore, it underpins the supramolecular and nano-material self-assembly and therefore has crucial consequences in material design across size scales.[6,7]

Considering the importance of NCIs, it is no wonder that a plethora of computational tools have been developed to visualize, calculate, probe, and predict them in systems of chemical and biochemical interest.[8,9] Depending on the scale of the complex studied, and the size of computing resources to be devoted to the cause, a series of methods have been deployed. The cheapest include empirically and knowledge-based pairwise interatomic potentials (IAPs) which are combined with conformation-generating software (for instance, genetic or Monte Carlo algorithms[10,11]) to create a docking function to probe molecules' binding modes.[12–14] Should the IAPs be linked to a more physically sound function (for instance, a quasi-time integrator), we could pursue not only the interesting interaction modes but also the trajectories linking them together; this is most famously implemented as molecular dynamics (MD).[15] The great advantage of those methods is in being able to investigate systems of very large size, even including the protein-protein or whole membrane complex simulations.[16–18] Such calculations are made feasible by the use of computationally cheap strategies - fitted force fields relying on the Lennard-Jones potential and the Coulomb law to describe van der Waals and electrostatic interactions, respectively.[19,20] These rely on long-

known and tested relationships which capture the correlation but fall short of explaining their underlying quantum origin. Indeed, the NCIs are not merely a monolith, but the origin of the interactions varies from electrostatic to quantum origin, and the final strength of interactions could be down to a combination of often subtle effects.[21]

To account for this detail, many quantum mechanics approaches are possible, most famously the density functional theory (DFT) which calculates the energy of the molecules and complexes, and the interaction energy is found by the supermolecular approach where a difference between the complex and individual energies often needs to be supplemented by a counterpoise correction for the elimination of the basis set superposition error (BSSE).[22] DFT is by no means the only method, especially due to its inability to fully derive the electron exchange and correlation.[23,24] Wavefunction methods, like coupled cluster, or CASSCF incorporate a more detailed description of the interactions, though pay for this accuracy with an exponentially large computational cost.[25] The chasm between accuracy and expenditure has been populated by a range of interesting, unique and often powerful algorithms, for example, energy decomposition schemes,[26] or the symmetry adapted perturbation theory (SAPT).[27,28] The topic of this paper, the NCI approach is an example of a similar but distinct class of real-space approaches, spearheaded by QTAIM[29] which aim to detect NCIs using physically sound and interpretable tools. In this respect it is worth mentioning the work of Espinosa *et. al.*,[30] who provided correlations for families of hydrogen bonds.

In order to go beyond classification of non-covalent families, efforts have been devoted within the NCI approach.[8,31] This approach is almost 14 years old and has become a widespread technique for qualitative analysis of NCIs due to its ease of use, speed, and graphic visualization of sites of interest.[32] It identifies the low-density low-reduced density gradient regions around topological -3 critical points which have been found to correspond to NCIs in real space as disc-shaped volumes. This approach makes use of partitioning of the interaction into intra- and interatomic components and even further into types of NCIs, for example, hydrogen bonding, repulsive contacts or van der Waals interactions.[32] The tech-

nique has been encoded as an easy-to-use program[33] and is available as a web tool,[34] which are, nevertheless, subject to an array of parameters which dictate the search for the region which could be integrated to give a further context. The integrand over the volume - the NCI index - is a promising avenue to adding a quantitative dimension to the NCI approach.[35,36] Although it allowed to go beyond families of NCIs, it has not been studied systematically to date (e.g. not applied to other test sets than the training); hence, the following article aims at exploring both the effect of NCI parameters on the found regions and an investigation of the relationship between the NCI index and interaction energy of the revealed interactions.

2 Theory

2.1 NCI Index

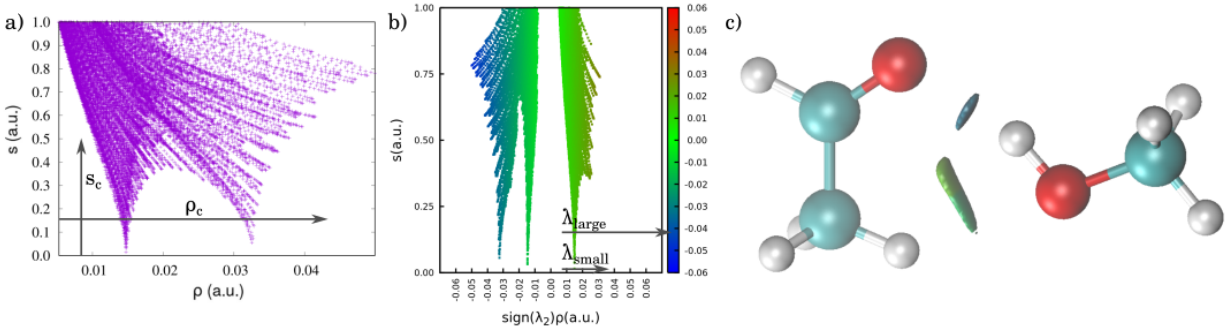

Figure 1: a) The RDG s_r - electron density ρ plot for an example interaction, b) the RDG s_r - signed electron density $sign\lambda_2\rho(\mathbf{r})$ plot for an example complex, c) the visualization of an example complex with the NCI region. There are four parameters that dictate the size of the NCI region which are depicted as quantities with arrows indicating their definition.

The NCI approach detects NCI regions as characterized by low electron density and low reduced electron density (RDG), $s(r)$

$$s(\mathbf{r}) = \frac{|\nabla\rho(\mathbf{r})|}{2(3\pi^2)^{1/3}\rho(\mathbf{r})^{4/3}} \tag{1}$$

and these so-defined low values are parameter-dependent. The latter, especially, is a

quantity highlighting the inhomogeneity of electron density that has been the cornerstone of the NCI method due to a pattern of troughs which can be tracked down as non-covalent interactions occurring in the complex (see Figure 1 a)) where through in bottom left corner correspond to NCIs). When plotted, these low-density regions correspond to volumes in real space located at the sites of interactions (see Figure 1 c)), and hence have been widely used to track the qualitative presence of interactions.[37] Further characterization is possible when utilizing the electron density Hessian to describe the nature of the troughs, more specifically the sign of its second eigenvalue, λ_2; the resultant scale dictates the type of interaction taking place: hydrogen bonds are strongly negative, repulsion - strongly positive, and weaker interactions are found with values in between (see Figure 1 b)). The second eigenvalue of the Hessian was chosen due to its ability to track density concentration or depletion - as expected from bonding or anti-bonding situations.[38,39]

The limits $sign\lambda_2\rho(\mathbf{r})$ for each interaction type were dependent on two additional parameters which partition the integration volume (see Figure 1 b)). If we want to consider intermolecular NCIs preferentially, we ought to consider one more parameter, bringing the total to five. This other parameter is the intermolecularity fraction, γ_{ref}, and it ensures that the density considered has significant contributions from both molecules in the complex. Hence, the density to be considered, the "integration domain", Ω_{NCI} must satisfy the following conditions

$$\begin{cases} \rho(\mathbf{r}_i) < \rho_c \\ s(\mathbf{r}_i) < s_c \\ \rho_A(\mathbf{r}_i) < \gamma_{ref}\rho(\mathbf{r}_i) \\ \rho_B(\mathbf{r}_i) < \gamma_{ref}\rho(\mathbf{r}_i) \end{cases} \tag{2}$$

and gets split into the three integration ranges

$$\begin{cases} \Omega_{Hydrogen\_bond} \; if \; sign(\lambda_2\rho(\mathbf{r}_i) < -\lambda_{large} \\ \Omega_{Van\_der\_Waals} \; if \; sign(-\lambda_{small} < \lambda_2\rho(\mathbf{r}_i) < \lambda_{small} \\ \Omega_{Repulsion\_contact} \; if \; sign(\lambda_2\rho(\mathbf{r}_i) > \lambda_{large} \end{cases} \tag{3}$$

An integral called the NCI index is taken over each of the integration domains

$$I_{n,X} = \int_{\Omega_X} d\mathbf{r}\rho^n(\mathbf{r}) \tag{4}$$

for X = {Hydrogen_bond, Van_der_Waals, Repulsion_contact} and where we have added an n$^{\text{th}}$ power dependency as it was not obvious to us why only the $n = 1$ case should be investigated, and the NCIPLOT software already calculated a series integrals with powers of $n = \{1, 4/3, 1.5, 5/3, 2, 2.5, 3\}$.

The NCIPLOT software requires an input of geometry of the complex and its wavefunction, but its underlying computation could vary depending on whether the input is from an independent SCF calculation or, alternatively, whether it has been constructed from a set of pre-calculated promolecular densities;[33] the latter is an option to speed up the calculation and enable the investigation of supramolecular and biomolecular systems. The features of the RDG remain stable across the exact density calculation detail, though, as we shall present, the integration of the density and the resultant NCI indices vary with the chosen density calculation method.

2.2 Symbolic Regression

It has been established in many previous publications[32,35,40] that there exists a relationship between the NCI index and the interaction energies as calculated by CCSD(T)/CBS or its approximations. The explicit relation is not easily derivable, nor is its closed form obvious.

For that reason, we decided to identify a symbolic relationship first and seek an explanation of the form second. Such an investigation is made possible with PySR - a symbolic regression library,[41] that has been used in such a manner in a variety of fields already.[42] The overall advantage and reason for using symbolic regression is in the interpretability of the resultant model, and the ability to probe the importance of various factors within it so that the relationship and physics behind the equations terms might be investigated.

PySR is high-performance and open-source with an efficient Julia backend and works very well on low-dimensional datasets where there are few qualifiers that describe a data point.[41] The library uses a multi-population evolutionary algorithm with adjustable mutation and cross-over rates to grow an equation and improve it through the evolution of new terms as offspring. It does additionally contain a few modifications (e.g. simulated annealing at lower temperatures) to both increase the diversity of investigated space of equations and to favor the fittest equations. PySR, therefore, stands out among other symbolic regression tools which are usually designed to handle more straightforward cases and therefore to be less scalable. It contains a simplification step to appreciate that the goal is to find the best equations hence balancing their complexity and accuracy.[41] It also allows us to tune the complexity of the equations sought by specifying the operators used and their complexity in the equation, hence preventing the search for overly complicated and fitted forms. We decided to restrict the equations to only containing addition, subtraction, multiplication, and simple powers, square and cube roots; the last three as a counterbalance to n^{th} powers in the NCI indices.

Nevertheless, the nature of the algorithm is inherently stochastic, and the path taken to construct the equation differs with every run, much like it does in other machine learning (ML) algorithms. In fact, the library is written much to mirror ML models whereby the model is trained on a subset (train) dataset and the performance of the resulting equations is evaluated on the data not used - the test set.

We also decided on a comparison with another ML technique to examine whether it is

possible to obtain a better albeit less interpretable model with the available data. For this purpose, we chose the very popular gradient boosting regression (GBR)[43] as implemented in the scikit-learn library.[44] It is an incredibly powerful algorithm that creates a strong model by combining weak learners into a model that is capable of detecting non-linear relationships as well as outliers, and able to find relationships with a limited amount of data, in hundreds and not millions of data points.[45]

3 Calculations

3.1 Training on the NCIA

In order to systematically study the non-covalent interactions, we turned to Rezac *et. al.'s* Non-Covalent Interactions Atlas (NCIA)[46] which contains large and diverse datasets for various types of non-covalent interactions along with very accurate reference energies as calculated with CCSD(T)/CBS. We decided to focus on the two most pervasive types of interactions first - on hydrogen bonds and dispersion interactions and, hence, used the HB375 and D1200 datasets,[47,48] and as suggested by their names, they contain 375 hydrogen bond and 1200 dispersion complexes, respectively. These datasets contain small, organic and inorganic dimeric complexes that were geometry-optimized to give an equilibrium separation. For each of the monomers in each of the complexes, we calculated the electron density using DFT with frozen core, PBE0-D3 functional, and def2-SVP basis set using Psi4[49] and converted the result to the wavefunction file type (.wfn) using Multiwfn 3.8.[50] We additionally calculated the electron density using the promolecular approach as present internally in the NCIPLOT 4.2 software.[33] As we described in the Theory section, an NCIPLOT calculation is subject to 5 parameters, hence we repeated the calculation for a five-dimensional set of parameters, namely:

1. λ_{large}: {0.2, 0.1, 0.07}

2. λ_{small}: {0.02, 0.015, 0.01}

3. ρ_c: {0.07, 0.05}

4. s_c: {1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3}

5. γ_{ref}: {0.95, 0.85, 0.75}

For each complex calculation, the NCI indices for the Hydrogen bond, van der Waals and Repulsion regions were calculated for $n = \{1, 4/3, 1.5, 5/3, 2, 2.5, 3\}$. We also investigated the effect of $n = 0$ - which corresponds to volume rather than charge - see SI for more details. We decided that the best set of parameters from the 432 possibilities would do the best job of predicting the energy of the intermolecular interactions versus the CCSD(T) reference. We could have looked at the simple correlation between the NCI indices at a given n and reference, but this approach would fall short should the equation be any more complicated than a simple linear relationship (also see Figure 3 c) and f)). For that reason, for each parametrised result, we carried out a search for the best symbolic regression equation. Each parametrised result gives five to nine 'hall of fame' equations for a given increasing complexity metric, and we decided to select the most accurate (and hence the most complex) equation as the representation of the performance of the parameter set. This allows for a calculation of not only the R^2 correlation coefficient between the NCI indices and energies but also of the mean absolute error (MAE, in kJ/mol throughout) to quantize the performance further. Each symbolic regression was performed with a 2:1 train:test split and the reported numbers correspond to the test set results. The train:test split was the same for every regression calculation (seeded with the name number). Below, we present the method and results using the promolecular approach, but a parallel analysis had been performed for the DFT densities; these results are compiled in the Supplementary Information.

Figure 2 shows an example heat map cross-section through a two-dimensional plane of the parameter space for both R^2 and MAE for the HB375 dataset. Such heat maps were possible through other cross-sections, and those are compiled in the Supplementary Information. The

starred square in Figure 2 shows the consensus set of parameters that minimizes MAEs for HB375 and D1200 datasets.

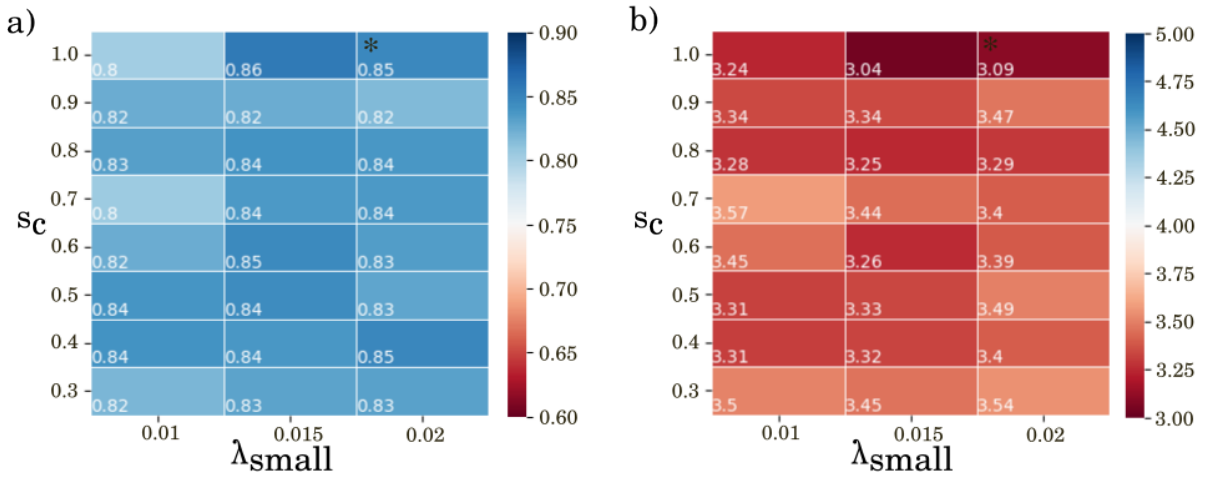

Figure 2: Heat map of a) R^2 correlation coefficient, b) mean absolute error (MAE) for two varying parameters: s_c and λ_{small} at set $\lambda_{large} = 0.2$, $\rho_c = 0.05$, and $\gamma_{ref} = 0.85$ for the HB375 dataset. The color maps are present to the right of the graph, and note that the best results have lowest and reddest MAE and largest and bluest R^2. The starred square with parameters $s_c = 1.0$ and $\lambda_{small} = 0.02$ is the consensus best set of parameters, also for the D1200 - for details, see the Supplementary Information. All underlying NCI indices calculations were carried out using the promolecular approach.

We decided that the lowest MAE is the ultimate criterion for the best set of parameters, along with the highest R^2 and simplest form of the equation. The parameters are: $\lambda_{large} = 0.2$, $\lambda_{small} = 0.02$, $\rho_c = 0.05$, $s_c = 1.0$, and $\gamma_{ref} = 0.85$. This best result corresponds to the following equations for HB375 - (5) and D1200 - (6). Equation (6) only depends on the van der Waals NCI indices, as other NCI indices were evaluated to zero in 80% of the D1200 complexes.

$$E_{Hydrogen\_bond}(\rho) = -(2.8 \times 10^3 \, (I_{2,van\_der\_Waals} + I_{2,Hydrogen\_bond}) + 2.7 \times 10^1 \sqrt[3]{I_{4/3,Hydrogen\_bond}}) \quad (5)$$

$$E_{van\_der\_Waals}(\rho) = -\left(5.0 \times 10^{1}\sqrt{I_{1,van\_der\_Waals}} - 7.3 \times 10^{1}\sqrt[3]{I_{5/3,van\_der\_Waals}}\right) \quad (6)$$

An analogous analysis was performed with the gradient boosting regression (GBR) algorithm. The presented performance is an average of three runs of the GBR code, as a non-negligible variance has been detected in individual runs, as this code is likewise non-deterministic. It gives the best performance for the above-mentioned set of parameters, similarly to PySR, with the HB375's $R^2 = 0.82$ and MAE = 3.35 kJ/mol and D1200's $R^2 = 0.75$ and MAE = 2.22 kJ/mol. This not only reinforces the choice of the parameters but also reveals the extent to which the energy results can be explained with the available density data. Figure 3 shows the scatter plots calculated versus the reference energy for the two methods, and also for the simple sum $-(I_{1,Hydrogen\_bond} + I_{1,van\_der\_Waals})$ against reference energy (notice the change in y-axis scale) to further showcase the need for the more complex form of the equation to appropriately capture the underlying behavior.

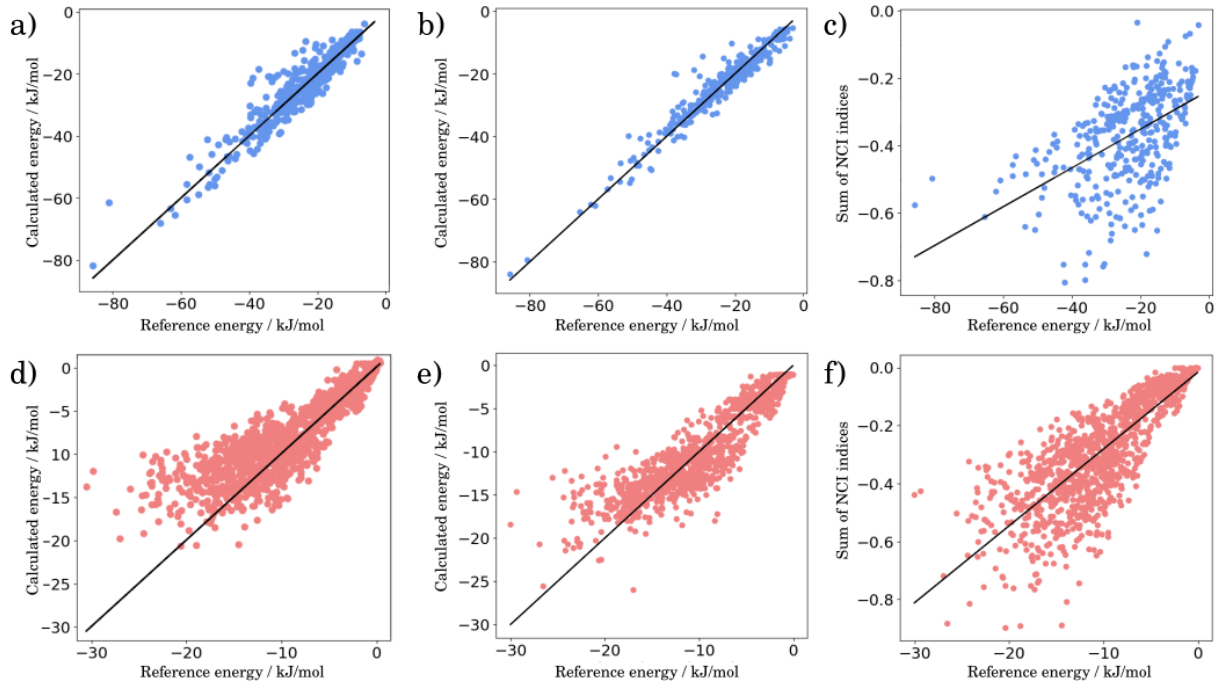

Figure 3: Scatter plot of calculated energy versus CCSD reference when the calculation is performed by a) PySR for the HB375 dataset, b) GBR for the HB375 dataset, c) Sum of $-I_{1,X}$ for the HB375 dataset, d) PySR for the D1200 dataset, e) GBR for the D1200 dataset, f) Sum of $-I_{1,X}$ for the D1200 dataset. All the graphs come from calculations with promolecular densities and using the optimum set of parameters.

Now, the choice of dataset and their analysis so far has been focussed on finding an equation for each of the types of interactions separately. It would be highly desirable to obtain a single unified equation that describes both of the interaction types simultaneously. Their mere sum overestimated the energy for almost all complexes (see Figure 4 a)); however, if we noted the I_{1,van_der_Waals} term in equation (5) was responsible for the calculation of the dispersion contribution for the complexes in the HB375 dataset and, therefore if we substituted this term by equation (6), we would bring about equation (7). This single composite equation performed much better as it avoided double-counting of the contributions to interactions (see Figure 4 b)).

$$E_{Int}(\rho) = -(2.8 \times 10^3 I_{2,Hydrogen\_bond} + 2.7 \times 10^1 \sqrt[3]{I_{4/3,Hydrogen\_bond}} \\ +5.0 \times 10^1 \sqrt{I_{1,van\_der\_Waals}} - 7.3 \times 10^1 \sqrt[3]{I_{5/3,van\_der\_Waals}}) \tag{7}$$

It should here be noted that we understand the dependence on energy of larger values for higher powers of NCI indices for hydrogen bonding akin to the relevance of charge for this type of interactions. Likewise, the dependence of the final equation on lower NCI indices for dispersion interactions could be due to the higher relevance of atomic volumes and contacts.

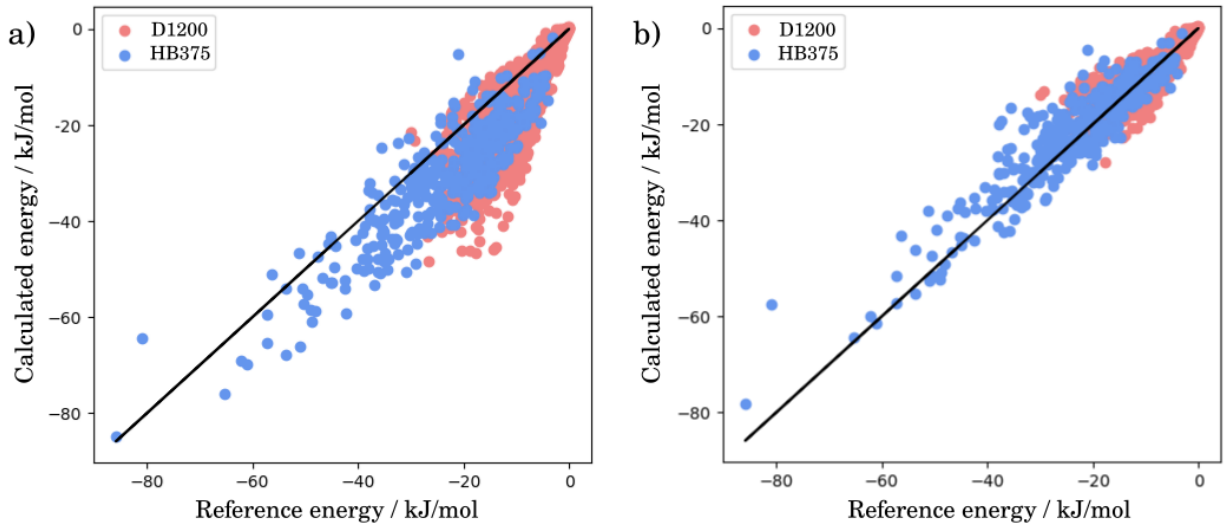

Figure 4: Scatter plot of calculated energy using a single equation versus CCSD reference for the color-coded datasets: blue is HB375 and pink is D1200. a) represents the exact sum of equations (5) and (6), and b) represents equation (7). The overall R^2 and MAEs are: a) 0.21 and 6.52 kJ/mol, and b) 0.86 and 2.63 kJ/mol.

There was perhaps a strange curving of the D1200 predictions which appeared persistent in all models, which was perhaps caused by the square and cubic roots in equations describing dispersion which was not the case for the dispersion contribution in the HB375 dataset. It was similarly seen in the GBR result, but not in the sum of NCI indices (Figure 3 f)). In trying to get a better estimate, we tried restricting the use of roots in sought-for equations in PySR, yet the obtained models worsened all following metrics. However, the prediction of D1200 energies using equation (7) significantly minimizes this perceived curving. For that

reason, we retained equation (5) as the best possible model for dispersion interactions.

Therefore, we have obtained a predictor equation that calculates the interaction energies as sub-chemical accuracy level (2.63 kJ/mol, 0.63 kcal/mol) even when using a rather approximate promolecular construction of electron densities. Using DFT densities produced more accurate estimates of MAE (2.20 kJ/mol, 0.53 kcal/mol) with similar correlation coefficients (0.86 versus 0.89), see Supplementary Information for the parallel analysis. It has also shown that this result is irrespective of the functional used to calculate the electron densities.

3.2 Testing on S66 dataset

Armed with the unique equation to describe a variety of non-covalent interactions, we turn our attention to a dataset of more complex and varied intermolecular interactions to test its accuracy in a new setting. The S66 dataset[51] was treated similarly with the densities calculated by both DFT (with the same functional/basis set) and using promolecular densities and the NCI indices calculated using the optimized parameters as identified in the previous section. The energies calculated using DFT densities gave the result in Figure 5 a), and ones found with equation (7) produced Figure 5 b).

The dataset is primarily split into complexes with binding energies below 40 kJ/mol which are overall well predicted, and 5 complexes with very strong interactions, which were all underpredicted by about 5 - 20 kJ/mol, depending on the origin of the electron densities. The strongly bound cyclic double hydrogen bonds (for example, as found in an acetic acid dimer) have been present in the HB375, and have similarly been underpredicted suggesting this to be a shortcoming of the promolecular approach, which perhaps did not appreciate the significance of the distortion of electron density by these strong hydrogen bonds. The remaining significant outliers were both uracil complexes (AcNH2-uracil dimer and uracil-uracil π stack).

When the strongly-bound hydrogen bond complexes were excluded the predictions gave

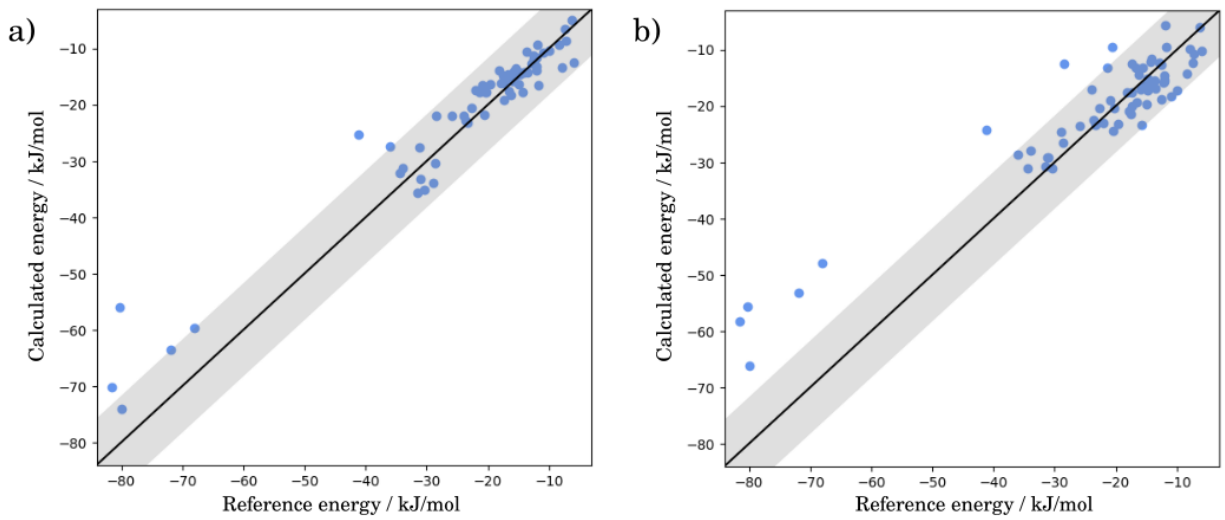

Figure 5: Scatter plot for the S66 dataset of the calculated energy using a single equation versus CCSD(T) reference using the a) DFT densities, and b) the promolecular densities. The overall R^2 and MAEs are: a) 0.92 and 3.18 kJ/mol, and b) 0.82 and 4.9 kJ/mol.

R^2 of 0.80 and MAE of 2.48 kJ/mol for the DFT-density approach and R^2 of 0.62 and MAE of 3.67 kJ/mol for the promolecular approach.

3.3 Limitations

Obtaining a symbolic relationship between the NCI indices and the binding energies of complexes resulted in simple equations capturing the underlying correlation. It also highlighted the non-equal contribution of the NCI indices representing the various types of non-covalent interactions. However, it should be noted that it did not produce a term describing the repulsive contribution to binding. In order to verify the lack of this term, we also analyzed the HB375x10 dataset, which contains compressed geometries (see Section 5 in the Supplementary Information). As the systems were compressed, energies were consistently overpredicted. If, on the contrary, we examine complexes with stretched interactions, a decrease in NCI indices is starker than expected, and the energy prediction - though still correlated - gives slightly lower energies than the reference (also see Figure 25 in SI for compressed geometry energy predictions of HB375). Therefore, the equations presented here work well

with equilibrium systems but should not be used to study out-of-equilibrium geometries.

It should also be noted that the PySR algorithm being non-deterministic. Hence, the equations presented here, though, found optimal by the algorithm run, might not be the absolute best equations that could exist to explain the relationship. When performing the same search for the symbolic equation, the final result usually changes the multiplicative constants slightly, and on several occasions, the exponents of the NCI index also differed slightly (e.g. between 4/3 and 1.5). Some examples of these alternative equations are given in SI.

4 Conclusions and Perspectives

We performed a detailed systematic study of the NCI method, evaluating the effect of its key parameters and determining the best values for these, therefore improving the definition of the method. These parameters corresponded to effectively considering the entire possible non-covalent region (not just the bond critical point). At the same time, we have investigated the relationship between the NCI indices obtained by NCIPLOT with the CCSD(T) reference energies for the HB375 and D1200 datasets to look for the symbolic equation linking the two. We have shown that a single equation can be created to describe both hydrogen bonding and dispersion interactions, which predicts the interaction energy for the combined HB375+D1200 dataset with a mean absolute error of well below chemical accuracy (using DFT: 2.20 kJ/mol, using promolecular densities: 2.63 kJ/mol). In order to verify the applicability of this equation, it was also tested on a different popular dataset S66 to evaluate how well it performs on new and more complicated complexes. The resultant predicted energies were still correct and very well correlated (R^2 = 0.92 and MAE = 3.2 kJ/mol) even with rather simplistic promolecular energies (R^2 = 0.82 and MAE = 4.9 kJ/mol). We should recall here that for promolecular densities, no SCF calculations were carried out. Hence, the quantitative NCI approach is therefore a promising alternative to (semi-)empirical functions

that calculate the non-covalent interaction energy with good accuracy even from the geometry. Probably also related to the use of promolecular densities, we were not able to capture correctly very strongly attractive neither repulsive interactions. Thenceforth, this approach can only be used for equilibrium geometries, yet we hope that this first step will nevertheless already provide the users with a useful tool for examining NCIs within complexes, and optimized conformers or ligand-protein interactions. Work is in progress for describing repulsive interactions which feature prominently in out-of-equilibrium conformations, and therefore are a key type of interaction to consider to extend the model to capture NCIs in all geometries.

Acknowledgement

The authors thank ANR TcPredictor S22JRAR036, ANR Fisciency S23JRAR060 and Emergence-SU H2Ox S23JR31014 for funding.

Associated Content

Supplementary Information is available free of charge at .... It contains the systematic evaluation of NCIPLOT parameter effects using both promolecular and DFT densities, and PySR and GBR models (sections 1 and 2), as well as specifies the details of the models used. Section 3 concerns the HB375 internal control of non-hydrogen bonding compounds. Section 4 shows the parallel analysis of the integral values and resultant symbolic energy equations using DFT-derived electron densities. Section 5 explores the HB375x10 geometries and the generalizability of the equations beyond equilibrium. Section 6 investigates the regression analysis of subsets of HB375 and D1200 datasets to ascertain the generalizability of the approach across NCI families.

References

(1) Kollman, P. *Chapter 2 Non-covalent forces of importance in biochemistry*; Elsevier, 1984.

(2) Sliwowski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Pharmacological Reviews. **2014**, *66*, 334–395.

(3) Dec, J.; Bollag, J.-M. Determination of Covalent and Noncovalent Binding Interactions between Xenobiotic Chemical and Soil. *Soil Science* **1997**, *162*, 88–874.

(4) Long, Y.; Hui, J. F.; Wang, P. P.; Xiang, G. L.; Xu, B.; Hu, S.; Zhu, W. C.; Lü, X. Q.; Zhuang, J.; Wang, X. Hydrogen bond nanoscale networks showing switchable transport performance. *Scientific Reports* **2012**, *2*.

(5) Hentschel, J.; Kushner, A. M.; Ziller, J.; Guan, Z. Self-Healing Supramolecular Block Copolymers. *Angewandte Chemie International Edition* **2012**, *51*, 10561–10565.

(6) Troselj, P.; Bolgar, P.; Ballester, P.; Hunter, C. A. High-Fidelity Sequence-Selective Duplex Formation by Recognition-Encoded Melamine Oligomers. *Journal of the American Chemical Society* **2021**, *143*, 8669–8678.

(7) Keinan, S.; Ratner, M. A.; Marks, T. J. Molecular zippers-designing a supramolecular system. *Chemical Physics Letters* **2004**, *392*, 291–296.

(8) Johnson, E. R.; Keinan, S.; Mori-Sánchez, P.; Contreras-García, J.; Cohen, A. J.; Yang, W. Revealing noncovalent interactions. *Journal of the American Chemical Society* **2010**, *132*, 6498–6506.

(9) Storer, M. C.; Zator, K. J.; Reynolds, D. P.; Hunter, C. A. An atomic surface site interaction point description of non-covalent interactions. *Chemical Science* **2024**, *24*, 160–170.

(10) Harrison, R. L.; Granja, C.; Leroy, C. Introduction to Monte Carlo Simulation. *AIP Conference Proceedings* **2010**, *1204*, 17–21.

(11) Paquet, E.; Viktor, H. L. Molecular Dynamics, Monte Carlo Simulations, and Langevin Dynamics: A Computational Review. *BioMed Research International* **2015**, 2314–6133.

(12) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics* **2002**, *47*, 409–443.

(13) Mooij, W. T. M.; Verdonk, M. L. General and targeted statistical potentials for protein-ligand interactions. *Proteins: Structure, Function and Genetics* **2005**, *61*, 272–287.

(14) Nataraj, P.; Khajamohiddin, S.; ; Jack, T. Software for molecular docking: a review. *Biophysical Reviews* **2017**, 91–102.

(15) Rapaport, D. C. *The Art of Molecular Dynamics Simulation*, 2nd ed.; Cambridge University Press, 2004.

(16) Elcock, A. H.; Sept, D.; McCammon, J. A. Computer Simulation of ProteinProtein Interactions. *The Journal of Physical Chemistry B* **2001**, *105*, 1504–1518.

(17) Erik, L.; S.P., S. M. Membrane proteins: molecular dynamics simulations. *Current Opinion in Structural Biology* **2008**, *18*, 425–431.

(18) Lewis, B. R.; Uddin, M. R.; Moniruzzaman, M.; Kuo, K. M.; Higgins, A. J.; Shah, L. M. N.; Sobott, F.; Parks, J. M.; Hammerschmid, D.; Gumbart, J. C.; Zgurskaya, H. I.; Reading, E. Conformational restriction shapes the inhibition of a multidrug efflux adaptor protein. *Nature Communications* **2023**, *14*, 3900.

(19) Lennard-Jones, J. E. On the Determination of Molecular Fields. II. From the Equation of State of a Gas. *Proceedings of the Royal Society* **1924**,

(20) Coulomb, C.-A. d. First dissertation on electricity and magnetism. *History of the Royal Academy of Sciences* **1785**,

(21) Elmi, C. S. L., Alex Quantifying Interactions and Solvent Effects Using Molecular Balances and Model Complexes. *Accounts of Chemical Research* **2021**, *54*, 92103.

(22) Helgaker, T.; Ruden, T. A.; Jørgensen, P.; Olsen, J.; Klopper, W. A priori calculation of molecular properties to chemical accuracy. *Journal of Physical Organic Chemistry* **2004**, *17*, 913–933.

(23) Kohn, W.; Sham, L. J. Self-consistent equations including exchange and correlation effects. *Physical Review* **1965**, *140*, A1133.

(24) Parr, R. G.; Weitao, Y. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press, 1995.

(25) Kümmel, H. Origins of the Coupled Cluster Method. *Theoretica Chimica Acta* **1991**, 81–89.

(26) Zhao, L.; von Hopffgarten, M.; Andrada, D. M.; Frenking, G. Energy decomposition analysis. *WIREs Computuational Molecular Science* **2018**, *8*, e1345.

(27) Szalewicz, K.; Jeziorski, B. Symmetry-adapted double-perturbation analysis of intramolecular correlation effects in weak intermolecular interactions: The He-He interaction. *Molecular Physics* **1979**, *38*, 191–208.

(28) Szalewicz, K. Symmetry-adapted perturbation theory of intermolecular forces. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 254–272.

(29) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory.International Series of Monographs on Chemistry*; Oxford Science Publications, 1990.

(30) Mata, I.; Alkorta, I.; Molins, E.; Espinosa, E. Universal Features of the Electron Density Distribution in Hydrogen-Bonding Regions: A Comprehensive Study Involving HX

(X=H, C, N, O, F, S, Cl,) Interactions. *Chemistry – A European Journal* **2010**, *16*, 2442 – 2452.

(31) Contreras-García, J.; Johnson, E. R.; Keinan, S.; Chaudret, R.; Piquemal, J. P.; Beratan, D. N.; Yang, W. NCIPLOT: a program for plotting non-covalent interaction regions. *Journal of Chemical Theory and Computation* **2011**, *7*, 625–632.

(32) Laplaza, R.; Peccati, F.; Boto, R. A.; Quan, C.; Carbone, A.; Piquemal, J.; Maday, Y.; Contreras-García, J. NCIPLOT and the analysis of noncovalent interactions using the reduced density gradient. *WIREs Computuational Molecular Science* **2021**, *11*, e1497.

(33) Boto, R. A.; Peccati, F.; Laplaza, R.; Quan, C.; Carbone, A.; Piquemal, J. P.; Maday, Y.; Contreras-García, J. NCIPLOT4: Fast, Robust, and Quantitative Analysis of Noncovalent Interactions. *Journal of chemical theory and computation* **2020**, *16*, 4150–4158.

(34) Novoa, T.; Laplaza, R.; Peccati, F.; Fuster, F.; Contreras-García, J. The NCIWEB Server: A Novel Implementation of the Noncovalent Interactions Index for Biomolecular Systems. *Journal of Chemical Information and Modeling* **2023**, *63*, 4483–4489.

(35) Wieduwilt, E. K.; Boto, R. A.; Macetti, G.; Laplaza, R.; Contreras-García, J.; Genoni, A. Extracting Quantitative Information at Quantum Mechanical Level from Noncovalent Interaction Index Analyses. *Journal of Chemical Theory and Computation* **2023**, *19*, 1063–1079.

(36) Peccati, F.; Desmedt, E.; Contreras-Garcia, J. A Regression Approach to Accurate Interaction Energies Using Topological Descriptors. *Computational and Theoretical Chemistry* **2019**, *1159*.

(37) Gibbs, D. F. R. K. M., G. V.; Cox A Connection between Empirical Bond Strength and the Localization of the Electron Density at the Bond Critical Points of the SiO Bonds in Silicates. *J. Phys. Chem. A* **2004**, *108*, 7643–7645.

(38) Bader, H., R. F. W.; Essén The characterization of atomic interactions. *J. Chem. Phys.* **1984**, *80*, 1943–1960.

(39) Bader, R. F. W. A Bond Path: A Universal Indicator of Bonded Interactions. *J. Phys. Chem. A.* **1998**, *102*, 7314–7323.

(40) Novoa, T.; Peccati, F.; Alonso, M.; Arias-Olivares, D.; Bohorquez, H.; Contreras-García, J. *New Developments in the Non Covalent Interaction (NCI) Index*; Elsevier, 2023.

(41) Cranmer, M. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl. *arXiv* **2023**,

(42) Tonda, A. Review of PySR: high-performance symbolic regression in Python and Julia. *Genetic Programming and Evolvable Machines* **2025**, *26*, 7.

(43) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **2001**, *29*.

(44) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(45) Hepp, T.; Schmid, M.; Gefeller, O.; Waldmann, E.; Mayr, A. Approaches to Regularized Regression – A Comparison between Gradient Boosting and the Lasso. *Methods of Information in Medicine* **2016**, *55*, 422–430.

(46) Rezac, J. Non-Covalent Interactions Atlas Benchmark Data Sets: Hydrogen Bonding. *Journal of Chemical Theory and Computation* **2020**, *16*, 2355–2368.

(47) Rezac, J. Non-Covalent Interactions Atlas Benchmark Data Sets: Hydrogen Bonding. *ACS Applied Materials and Interfaces* **2020**, *16*, 2355–2368.

(48) Řezáč, J. Non-Covalent Interactions Atlas benchmark data sets 5: London dispersion in an extended chemical space. *Physical Chemistry Chemical Physics* **2022**, *24*, 14780–14793.

(49) Turney, J. M. et al. Psi4: An open-source ab initio electronic structure program. *WIREs Computational Molecular Science* **2011**, *2*.

(50) Lu, T. A comprehensive electron wavefunction analysis toolbox for chemists. *Journal Computational Chemistry* **2024**, *161*, 082503.

(51) Rezac, J.; Riley, K. E.; Hobza, P. S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *Journal of Chemical Theory and Computation* **2011**, *7*, 2427–2438.