Development and testing of new exchange correlation functionals

Dissertation for the degree of Doctor Philosophiae

Keld T. Lundgaard

January 2014

Center for atomic-scale materials design
Department of physics
Technical university of Denmark

Technical University of Denmark
Department of Physics
Center for Atomic-scale Materials Design
Building 311, DK-2800 Kongens Lyngby, Denmark
www.camd.dtu.dk

Preface

This thesis is submitted in candidacy for the Ph.D. degree in physics from the Technical University of Denmark (DTU). The work has been carried out from February 2011 to January 2014 at the Center for Atomic-scale Materials Design (CAMD), Department of Physics, DTU. Supervisors have been Professor Karsten W. Jacobsen, Director of CAMD, Kristian Thygesen, Associate Professor, and Thomas Bligaard, Senior Staff Scientist at the SUNCAT Center for Interface Science and Catalysis at SLAC National Accelerator Laboratory, Menlo Park, California. This thesis contains results obtained in collaboration with other researchers at CAMD and SUNCAT.

California, January 31, 2014
Keld T. Lundgaard

Abstract in English

Catalysts are used in 90% of the world's chemical processes to produce 60% of its chemical products, and they are thus very important to our modern society. We therefore seek to better understand current catalytic materials, so that we can find alternatives that will improve the energy efficiency, selectivity or similar of current chemical processes, or to make new technologies economical feasible.

Kohn-Sham density functional theory (KS-DFT) has proven to be a powerful theory to find trends in current catalytic materials, which can empower a more informed search for better alternatives. KS-DFT relies on accurate and efficient approximations to the exchange correlation functional, yet these functional approximations have lacked a systematic way to estimate the underlying uncertainties. A Bayesian error estimation approach provides a mechanism for calculating approximative uncertainties, and so accurate, computationally feasible exchange-correlation approximations that incorporate it have been called for.

This thesis presents significant steps forwards towards providing general applicable exchange-correlation functional approximations with Bayesian error estimation capabilities. A semi-empirical approach was used with a machine learning toolset to improve accuracy and transferability of the functional approximations. The toolset includes Tikhonov regularization of smoothness in a transformed model space, for ensuring sensible model solutions; an explicit model compromise with a geometric mean loss function, for ensuring generally applicable models; a robust MM-estimator loss function, for ensuring resistance to outliers in data; and a hierarchical bootstrap resampling estimating prediction error validation method, for selecting the model complexity that provide best transferability outside the training data.

Three new semi-empirical functional approximations have been made: BEEF-vdW, mBEEF, and mBEEF-vdW. It is shown that these functionals are able balance the accuracy of predicting energetics of covalent and non-covalent chemistry better than any comparative functional that we have tested, and they could therefore become the functional approximations of choice for understanding chemical processes at the solid-gas and solid-liquid interfaces.

Resumè på dansk

Katalysatorer anvendes i 90% af verdens kemiske processer til fremstilling af 60% af vores kemiske produkter, og de er derfor meget vigtige for vores moderne samfund. Vi søger derfor bedre at forstå anvendte katalytiske materialer, således at vi kan finde alternativer, der kan forbedre energieffektiviteten, selektivitet og lignende for aktuelle kemiske processer, eller for at gøre nye grønne teknologier økonomisk rentable. Kohn-Sham tæthedsfunktionalteori (KS-DFT) har vist sig at være en nyttig teori til at finde tendenser i kendte katalytiske materialer, og derved bidrage til en mere informeret søgning efter bedre alternativer. KS-DFT afhænger af nøjagtige og effektive tilnærmelser af exchange-korrelations funktionalet, og disse funktional approksimationer har manglet en systematisk måde at vurdere de underliggende usikkerheder på. Den Bayesiansk fejlestimering metode har er en måde hvorpå man kan approksimere disse usikkerheder, og det har skabt efterspørgsel efter nøjagtige og hurtige exchange-korrelations approksimationer med mulighed for disse estimationer. Denne afhandling præsenterer betydelige fremskridt i retning af at skabe generalle brugbare estimation exchange-korrelation funktionale approksimationer med Bayesian fejlestimering tilgængelige. En semi-empirisk metode blev brugt sammen med fitte teknikker til at forbedre nøjagtigheden og overførbare funktional approksimationer. Disse teknikker inkluderer Tikhonov regularisering af glathed i et transformeret model rum, for at sikre fornuftige overførbarhed; et specifikt model kompromis med et geometrisk gennemsnit kost funktion, for at sikre generelt anvendelige modeller; en robust MM-estimator kostfunktion, for at sikre modstandsdygtighed over for outliers i data; og en hierarkisk bootstrap resampling estimering af forudsigelses fejlene, til valg af model kompleksitet, der giver bedst omsættelighed udenfor træningsdata. Der blev skabt tre nye semi-empiriske funktionaler: BEEF-VDW , mBEEF og mBEEF-VDW. Det bliver vist, at disse funktionaler er i stand balancere nøjagtigheden i at forudsige energetik for kovalent og ikke-kovalent kemi bedre end nogen af de funktionel vi har sammenlignelig dem med, og de kan derfor blive de funktionelle approksimationer som man i fremtiden vil bruge for at forstå kemiske processer i faststof-gas og faststof-væske grænsefladerne.

Acknowledgements

My acknowledgment should go to many people for the successful completion of this work, and the following list will unfortunately not cover all those who should be mentioned.

To Karsten W. Jacobsen for guidance providing me the opportunity and encouragements to explore during my Ph.D. And for always be cheerful at our many meetings.

To Thomas Bligaard for having me spent much of my Ph.D. in California, with a belief that something good were to come out of my struggles. For many many long and fruitful discussions. For being a model for how to think.

To Jess Wellendorff. It has been great to be under your wing for these three years. Thank you for many great times during hard work, discussions, travels or in breaks between it all. And thank you for minimizing my time spent on non-converging DFT calculations. I have a hard time visualizing a successful Ph.D. without you - at least an as enjoyable one.

To co-authors of the manuscripts included in this thesis: Andreas Møgelhøj, Vivien Petzold, David D. Landis and Jens K. Nørskov.

To all my colleges at CAMD and SUNCAT for good conversations, collaboration and inspiring work. I would especially say thank you for all the people at SLAC who have been a part of creating a feeling being a part of a bigger family.

To the benefactors of my stay at Stanford: Danmark-Amerika fondet, Civilingeniør Frants Allings fond, Berg-Nielsen Legat, and Reinholdt W Jorck og Hustrus Fond.

To the administration at both CAMD/DTU and SUNCAT/SLAC for making everything possible without major administrative pain, and computer problems.

To Christian Perez for support and final proof reading of abstract.

To my family for always being supportive.

List of included papers

Paper I

Density Functionals for Surface Science: Exchange-correlation Model Development with Bayesian Error Estimation

J. Wellendorff, K. T. Lundgaard, A. Møgelhøj, V. Petzold, D. D. Landis, J. K. Nørskov, T. Bligaard, and K. W. Jacobsen Physical Review B 85, 235149 (2012)

Physical Review B 85, 235149 (2012)

Paper II

mBEEF: An accurate semi-local Bayesian error estimation density functional

Jess Wellendorff, Keld T. Lundgaard, Karsten W. Jacobsen, and Thomas Bligaard

Submitted to The Journal of Chemical Physics

Contents

Abstract in English 3

Resumè på dansk 5

Acknowledgements 7

1 *Introduction* 13

2 *The Kohn-Sham exchange-correlation functional approximation* 15

2.1 *Strong and weak bonding* 15

2.2 *KS-DFT* 16

2.3 *The five-rung ladder of density functionals* 18

2.4 *LSDA* 19

2.5 *Generalized Gradient approximation* 20

2.6 *MGGA* 22

2.7 *The non-local functionals for dispersion* 22

2.7.1 *Rutgers-Chalmers non-local correlation approximation* 23

2.7.2 *VV09 and VV10 non-local correlation approximations* 25

2.8 *Hyper-GGA and hybrids* 25

2.9 *Random Phase approximation + corrections* 26

2.10 *The self-interaction error* 26

2.11 *Solving the Kohn-Sham equations computationally efficiently* 28

3 *Datasets for training and validation* 31

3.1 *Datasets of materials properties* 31

3.1.1 *G3/99 and G2/97: Molecular formation energies* 31

3.2 *RE42: Molecular reaction energies* 32

3.3 *DBH24/08: Molecular reaction barriers* 32

3.4 *S22 and S22x5: Non-covalent bonding* 33

3.5 *Crystalline solids* 34

3.5.1 *Sol34Ec* 34

3.5.2 *Sol27* 34

3.5.3 *Extended solids dataset of Sol54Ec and Sol58LC* 35

3.6 *CE27: Chemisorption on solid surfaces* 35

4 *Machine learning methods* 37

4.1 *Parametrization* 37

4.2 *The Bayesian connection* 38

4.2.1 *Probability theory and Bayes' theorem* 38

4.2.2 *Maximum-likelihood model for model with Gaussian noise* 39

4.2.3 *Overfitting* 40

4.2.4 *Prior model expectation and regularization* 40

4.2.5 *The frequentist interpretation* 41

4.3 *Minimizing the cost function* 41

4.3.1 *The effective number of parameters* 43

4.3.2 *The Bayesian error estimation ensemble* 43

4.3.3 *Origo of the solution* 45

4.3.4 *Tikhonov regularization* 45

4.4 *Optimal model complexity.* 46

4.4.1 *Cross-validation* 46

4.4.2 *Leave-One-Out Cross-Validation* 47

4.4.3 *Bootstrap resampling and the Bootstrap .632 estimate* 48

4.5 *A geometric mean cost and loss function* 49

4.5.1 *Geometric mean of cluster cost functions* 49

4.5.2 *Geometric mean of cluster loss functions* 51

4.5.3 *Cluster-LOOCV for model compromise* 52

4.5.4 *Hierarchical cluster-Bootstrap.632* 52

4.6 *Robust fitting* 54

4.6.1 *M-estimators* 54

4.6.2 *The iterative reweighting least squares (IRWLS)* 55

4.6.3 *The MM-estimator* 57

5 *Bayesian Error Estimation functionals* 59

5.1 *Inspiration for density functionals with error estimation* 59

5.2 *The first density functional with Bayesian error estimation* 60

5.3 *Investigation of the constraints in the GGA formalism* 63

5.4 *Full machine learning approach Bayesian density functional.* 64

5.4.1 *Datasets* 64

5.4.2 *Model space* 64

5.4.3 *Model selection with regularization and bootstrap resampling* 65

5.4.4 *Model compromise for multiple datasets* 66

5.4.5 *Discussion* 66

6 *The BEEF-vdW functional* 67

6.1 *The parametrized GGA+vdW model space* 68

6.2 *The exchange-correlation model selection* 69

6.2.1 *Fitting individual materials properties* 69

6.2.2 *The individually trained exchange-correlation models* 71

6.3 *Compromise model for several materials properties* 73

6.4 *The BEEF-vdW functional* 73

6.5 *The Bayesian error estimate* 75

6.6 *Assessment of BEEF-vdW functional* 76

6.7 *Summary and discussion* 77

6.7.1 *The strengths of the BEEF-vdW study* 77

6.7.2 *Limitations of the BEEF-vdW study* 78

7 A meta-GGA Bayesian Error Estimation Functional 81
7.1 The model space 81
7.2 Training datasets 82
7.3 Exchange model selection 83
7.4 The mBEEF functional 84
7.5 Benchmark 85
7.6 The model compromise investigated 85
7.7 Summary and discussion 86
7.7.1 Strengths 87
7.7.2 Limitations 87

8 Robust fitting of exchange correlation functionals 89
8.1 The model space 89
8.2 The training datasets 90
8.3 Model selection 91
8.3.1 Rescaling datasets from the Bayesian error estimation 91
8.3.2 What are outliers in the the dataset 91
8.3.3 Constants for the MM procedure 93
8.3.4 The initial estimator 94
8.3.5 Defining the loss function 94
8.4 Outlier detection in RE42 96
8.5 Assessment of best loss functions for the exchange correlation model selection 97
8.5.1 Comparing loss functions for the individual datasets 97
8.5.2 Benchmark 104
8.6 Summary 106

9 Future development 107
9.1 Model space 107
9.2 Training Datasets 108
9.3 Method development 108

10 Bibliography 109

1 Introduction

Catalysts are used in 90% of the world's chemical processes to produce 60% of its chemical products.[1] Catalysis can thus be thought of as the backbone of our the modern society. We therefore seek to better understand current catalytic materials, so that we can find alternatives that will improve the energy efficiency, selectivity or similar of current chemical processes, or to make new technologies economical feasible.

Today, computer models are used to augment the problem solving ability of human intelligence. Following Moore's law, the computational power has increased exponentially over several decades. Allowing simulations with Kohn-Sham Density functional theory (KS-DFT) for elucidating mechanisms and fundamental trends in enzymatic and heterogeneous catalysis, for designing chemically new active materials by "electronic structure engineering".

When studying the physical world itself every measurement taken by an apparatus will have an uncertainty/error associated with it. To properly compare measurements from different models we need an estimate of this measurement uncertainty. Likewise when using a computer to simulate nature, we should calculate an uncertainty estimation associated with every measurement taken on this virtual apparatus.

Density Functional Theory (DFT) is an exact electronic structure theory where the electrons density is used instead of the electron wavefunction, and it as proven successful within DFT to use the Kohn-Sham theory approach, where the interaction between the electrons is described by an exchange-correlation functional. The exact functional is unknown, but even if the exact functional were known, it would be computational intractable and require an approximation. For the last five decades or so, scientist have developed models of the exchange-correlation functional that could provide enough accuracy to gain insights into material science phenomena while still being computationally tractable. Two approaches have been used to develop new generally applicable models of the exchange correlation functionals: reductionism and the empiricism. The reductionist

[1] "Recognizing the Best in Innovation: Breakthrough Catalyst". R&D Magazine, September 2005, p. 20.

seeks to use model systems where exact properties are known, while empiricists use empirical data to fit the functional. Many proposed functionals of these two approaches, or the semi-empirical combination of the two, have provide accurate measurements yet none provided systematic way to estimate the uncertainty on the acquired result.

The development of Bayesian Error Estimation ensemble functionals begins here, prompted by the development of atomic pair potentials that use Bayesian statistics to provide an error estimates. In the Bayesian approach to statistics, a direct connection between the model and the data is proposed which includes a deliberate accounting of the prior knowledge. This allows one to ask "given the data what is the best model and what are the uncertainty of the model parameters?"

To make error estimations for an exchange correlation functional, an optimized model of the functional is needed. Semi-empirical model development has three pillars: model space, training data, and model selection. The model space determines to what extend local, semi-local or non-local information of the electron distribution is know. More information allows for a more complete model, but carries a higher computational cost. The training data are well known quantities from real materials or other models, and are selected to promote transferability of the model for systems outside the training data. Finally one selects a model within the defined model space that best captures the material properties of interest. Here one needs to take care not to overfit the data, e.g. making the model too complex within the given model space.

In this thesis my goal is to present an overview of insights that have been gained in developing exchange-correlation functionals with error estimations, and thereby put my scientific contributions into a proper context.

First will provide an introduction to density functional theory, an overview of the training data, and an introduction to the machine learning tools used throughout the studies. Secondly, I will go through previous work on Bayesian error estimation functionals, up until when I begun my work on the subject, followed by the three studies that I have contributed to. In the end, I will conclude the thesis with an outlook on possible fruitful future directions in Bayesian error estimation functionals development.

2 The Kohn-Sham exchange-correlation functional approximation

The following chapter a brief introduction to the Kohn-Sham approach to density functional theory (DFT). We will use Perdew's methaphorical ladder (Jacob's ladder) to create a hierarchy of Density functional approximations (DFAs), and introduce the zoo of publicized DFAs, that has become available to DFT users, but with an emphasis on the DFAs used in theoretical surface science. First however, we will take a view at the different chemistry such functionals need to describe.

2.1 Strong and weak bonding[1]

The types of inter-atomic bonding in matter are commonly grouped as; Ionic, covalent, metallic, van der Waals (vdW) like, and hydrogen bonds. The first three provide strong bonding interactions of the nuclei and the nearby electron density, and these are characterized by their, relative to the last two type, small bond lengths and large density overlap. These types of bonding are the main responsibility of making matter around us stable, and they counteract the repulsion due to ion-ion electrostatics (Coulomb) and the Pauli exclusion principle for short ion-ion distances.

The bonds in ionic crystals (e.g., solid NaCl) are formed due to the large difference in electronegativity between the ions, where charge transfer leads to approximately closed-shell ions and large electrostatic attractions between them. The covalent bonds, which are both present in molecules and solid crystals, are created from a redistribution of the electron density due to a hybridization between pairs of the valence electrons, which leads to new bonding and antibonding states. Covalent bonds can therefore be described as "charge sharing" between atoms. For the metallic bonds the electrons in the solid material are completely delocalized and thereby shared by the entire crystal. The positive ions are situated in a sea of shared electrons and the conducting electrons can be seen as a fermi gas of

[1] Based on Kittel [2005], Salam [2009], Martin [2004]

nearly free electrons.

The hydrogen bonding and the vdW interactions are most often a weaker interaction than the above, but for some systems they play a very important role in creating stability; this is for instance the case in biological matter (e.g. in proteins and DNA), for rare-gas chemistry, and for soft matter. In the hydrogen bond the attraction arises from an interaction between the two species in a link of the form $\mathrm{A} - \mathrm{H} \cdots \mathrm{B}$, where A and B are strongly electronegative atoms, such as F, O, and N. The interaction is therefore caused by an electrostatic interaction between the polarized hydrogen atom and atom B, where the electronegative host A is neutralizing much of the single electron of the hydrogen atom, thus making the interaction between H and B weak. These interaction energies are found to be in the order of 0.1 eV, which can be compared to the covalent bond between two hydrogen atoms with a bonding strength of 4.8 eV.

Lastly, the van der Waals forces, the weakest interaction of the five, but virtually always present. This interaction is even present in the limit of large separation between the interacting fragments with no overlap. The forces arise from spontaneous charge density fluctuations, that result in transient electromagnetic fields which induce dipole and multi-dipole interactions between distant charge densities. The correlation of these temporary fields leads to an attractive force between the charges. These forces are long-range but decay algebraically with the separation. For the considerations here, only the non-retarded regime is considered, where the length between the interacting charges are small enough so that finite speed of light plays little role; hence the response time of the dipoles is longer than the interaction time between the charges.

2.2 *KS-DFT*[2]

Density functional theory has today become a workhorse for electronic structure calculations. The Kohn-Sham density functional theory introduced practical estimates for the ground-state energy and electron density of the many-electron system. With proper approximations, to be introduced later, the methods is able to computationally affordably predict the sizes and shapes of molecules, the crystal structures of solids and the work required to stretch or break chemical bonds; thus covering all the bonding types introduced above.

The systems that we are interested in can be described by a Hamiltonian of the form

$$\hat{H} = \sum_{i=1}^{N} \left[-\frac{1}{2}\nabla_i^2 + v(\boldsymbol{r}_i) \right] + \frac{1}{2}\sum_i \sum_{i \neq j} \frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} + V_{nn}, \tag{2.1}$$

[2] This section is based on Perdew and Schmidt [2001], and is by no means intended to provide a excessive overview of Kohn-Sham theory. For details consult e.g. Perdew and Kurth [2003].

where V_{nn} is the electrostatic potential energy coming from interaction among the nuclei. The equations here are given in Hartree atomic units, hence $\hbar = e^2 = m = 1$. The external potential $v(\boldsymbol{r})$ is usually describing the interaction between the nuclei and the electrons, but can also include external fields from fragments not considered in the Hamiltonian. The ground-state eigenfunction of $\hat{H}$ is a correlated N-electron wavefunction, this function has 3N arguments, which makes it computationally problematic to deal with in terms of evaluating the Hamiltonian and storing the results for large systems. The Kohn-Sham density functional theory proves that instead of solving the problem for the interacting system, one can find the ground state densities $n_\sigma(\boldsymbol{r})$ ($\sigma =\uparrow$ or $\downarrow$ spin states) and energy E in principle exactly, by solving a self-consistent one-electron Schrödinger equation for N orbitals $\psi_{\alpha\sigma}(\boldsymbol{r})$, which each only are functions of the 3 spacial arguments $\boldsymbol{r} = (x, y, z)$. The self-consistent orbitals are in these equations implicit functionals of the electron density of the spin up and down states ($n_\uparrow(\boldsymbol{r})$ and $n_\downarrow(\boldsymbol{r})$). The Kohn-sham equations are

$$\left[-\frac{1}{2}\nabla^2 + v(\boldsymbol{r}) + \int d\boldsymbol{r}' \frac{n(\boldsymbol{r}')}{|\boldsymbol{r}' - \boldsymbol{r}|} + v_{xc}^{\sigma}(\boldsymbol{r})\right] \psi_{\alpha\sigma} = \epsilon_{\alpha\sigma}\psi_{\alpha\sigma}(\boldsymbol{r}), \qquad (2.2)$$

$$n(\boldsymbol{r}) = n_\uparrow(\boldsymbol{r}) + n_\downarrow(\boldsymbol{r}), \qquad (2.3)$$

$$n_\sigma(\boldsymbol{r}) = \sum_\alpha \Theta(\mu - \epsilon_{\alpha\sigma})|\psi_{\alpha\sigma}(\boldsymbol{r})|^2, \qquad (2.4)$$

$$E = \sum_\sigma \int d\boldsymbol{r}\tau_\sigma(\boldsymbol{r}) + \int d\boldsymbol{r} n(\boldsymbol{r})v(\boldsymbol{r}) + \frac{1}{2}\int d\boldsymbol{r} \int d\boldsymbol{r}' \frac{n(\boldsymbol{r})n(\boldsymbol{r}')}{|\boldsymbol{r}' - \boldsymbol{r}|} + E_{xc}[n_\uparrow, n_\downarrow] + V_{nn}, \qquad (2.5)$$

$$\tau_\sigma = \sum_\alpha \Theta(\mu - \epsilon_{\alpha\sigma})\frac{1}{2}|\nabla\psi_{\alpha\sigma}(\boldsymbol{r})|^2, \qquad (2.6)$$

$$v_{xc}^{\sigma} = \frac{\delta E_{xc}}{\delta n_\sigma(\boldsymbol{r})}. \qquad (2.7)$$

The chemical potential μ is in equations 2.4 and 2.6, so that $\int d\boldsymbol{r} n(\boldsymbol{r}) = N$. The fermion occupation numbers are derived from the step function $\Theta(x)$ as 0 for $x < 0$ and 1 for $x > 1$.

If the exchange-correlation energy E_{xc} is omitted along with its functional derivative $\delta E_{xc}/\delta n_\sigma(\boldsymbol{r})$ (the exchange-correlation potential), one ends up with the Hartree equations without the self-interaction correction. If instead the correlation energy is omitted in $E_{xc} = E_x + E_c$, but where the exchange energy E_x is treated exactly, the equations are that of the Hartree-Fock approximation.

Both the Hartree (which neglects E_{xc}) and the Hartree-Fock (neglects the correlation energy E_c and with the exact exchange) approximation fail dramatically in describing chemical bond energies compared to the simplest E_{xc} approximation.[3] Both the exchange and correlation energy are important for describing chemistry and the combined two can thus be said to be "nature's glue".

The exchange correlation energy is found by integrating exchange-correlation density per particle over all space:

$$E_{xc} = \sum_{\sigma=\uparrow\downarrow} \int \epsilon_{xc}\, n(\boldsymbol{r}) d\boldsymbol{r},$$

where $\epsilon_{xc} = \epsilon_x + \epsilon_c$ is the exchange correlation energy density that usually depends on the electron density $n(\boldsymbol{r})$, and other quantities, and what those quantities are is the topic of the following.

2.3 *The five-rung ladder of density functionals*

Jacob's ladder of density functional approximations for the exchange-correlation energy, is a systematic classification of the density functional approximations in DFT.[4] At each rung more complex and more global ingredients are added to the approximations, thus making it possible for the functional to provide a better approximation to the exact functional. This comes however with an added computational cost. The original Jacobs ladder consists of the following ordered from the least complex to the most complex: The local spin density approximation[5] (LSDA), the generalized-gradient approximation (GGA), meta-GGA (MGGA), hyper-GGA, and the random phase approximation.

Since the formulation of the ladder a number of functionals have however been formed that are using non-local density information. A rung between the the MGGA and Hyper GGA is therefore inserted, and the new ladder will be used to organize the functionals, see figure to right.

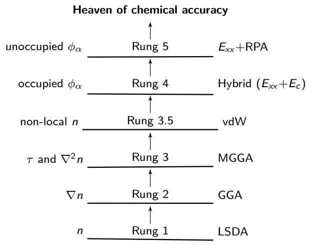


Figure 2.1: The revised Jacob's ladder with non-local density rung included.

Starting again from the lowest rung: The LDA uses only the local density as input, while at the second rung and the third rung the semi-local dependency of the density (GGA) and the KS orbitals (MGGA) are added. At the inserted rung 3.5 the non-local density is included, used in the van der Waals funtionals. The hyper-GGA rung introduce the nonlocal dependence of the occupied KS orbitals in the exact exchange energy density, and thereby only approximating the correlation energy. At the fifth-rung the unoccupied KS orbitals are added so that the correlation energy can be calculated through the Random phase approximation.

[3] See table in Perdew and Schmidt [2001].

[4] J. P. Perdew and K. Schmidt. Jacob's ladder of density functional approximations for the exchange-correlation energy.Perdew and Schmidt [2001]

[5] Kohn and Sham [1965]

To construct good approximations to the exact functional at each rung, two different approaches can be taken: That of the reductionist and that of the empiricist.

The strategy of the reductionist is to deduce the functional form from known constrains of the exact functional. He or she will use model systems where the exact functional is known exactly, such as the uniform electron gas (UEG); which is also known as the Homogeneous electron gas (HEG), and as jellium. At each level of complexity the reductionist will try to derive the simplest model that can take into account of the most relevant exact constraints, and first then test the functional on relevant data.

The empiricist, on the other hand, uses empirical evidence as reference data for a to fit a parametrized model of the functional. He will search for the simplest model that is able to properly reproduce the empirical data, and using tests to ensure that it is not an overfit. When the exact constraints to the function are also used, the approach is referred to as semi-empirical.

A challenge for the empirical or semi-empirical approach is to ensure that the model that is created is transferable to systems outside the training data, and that the model is thus not an overfit. Furthermore, reliable training datasets are needed for this approach. Which is why constraints are also taken into account in this approach. The reductionist on the other hand needs to make many choices about what constraints to be accounted for and how. These choices will in the end be made from looking at empirical data for verifying different models, or at insights at what functional forms comes out of the empirical approach. An interplay between the two methods are therefore used and one call talk of a synergetic relationship between the two approaches.

2.4 LSDA

The local spin density approximation (LSDA) by Kohn and Sham,[6] can be said to be the mother of all density functional approximations:

$$E_x^{LSD}[n_\uparrow, n_\downarrow] = \int d\boldsymbol{r} n(\boldsymbol{r}) \epsilon_{xc}^{unif}(n_\uparrow(\boldsymbol{r}), n_\downarrow(\boldsymbol{r})),$$

where $\epsilon_{xc}^{unif}(n_\uparrow, n_\downarrow)$ is the exchange-correlation energy density per particle for the uniform electron with spin densities $n_\uparrow$ and $n_\downarrow$. The UEG is a very important system for creating density functional approximations as it has a number of properties that can be found very accurately; the $\epsilon_{xc}^{unif}(n_\uparrow, n_\downarrow)$ is for instance accurately known and parametrized.

[6] Kohn and Sham [1965]

The LSDA has proven to be a surprisingly accurate approximation for solids and solid surfaces, where the electron density of the valence electrons is very homogeneous, and thus has a very big resemblance to the uniform electron gas. The explanation for why the approximation works for solids thus relies on that the approximation satisfy many exact constraints for the homogeneous electron gas.[7] An important observation for the functional is that it is almost always a better approximation for E_{xc} than it is for E_x or E_c separately. This is due to an observed error cancelation between the two terms in the approximation.

On the other hand it has been observed LSDA is an inadequate approximation to systems where the electron density varies significantly, e.g. for molecular bonds. The molecular geometries and vibration frequencies are reasonably described by LSDA, but thermochemistry predictions are far off and the intra-molecular covalent bond energies are vastly overestimated.[8] The LSD approximation was therefore not widely adopted in the quantum chemistry community.

2.5 Generalized Gradient approximation

For the next functional class the gradient of the electron density, ∇n, is added to the LSD information. This is most often done in the Generalized Gradient approximation (GGA) formalism. GGAs are therefore presented as a semi-local approximation.

In the GGA exchange energy formalism, an enhancement factor scales the local energy of the uniform electron gas, $\epsilon_x^{unif}(n)$, through an enhancement factor F_x, hence

$$\epsilon_x^{GGA}(n, \nabla n) = \epsilon_x^{unif}(n) F_x(s), \tag{2.8}$$

where s is the reduced density gradient defined as

$$s = \frac{|\nabla n|}{2k_F n} \in [0, \infty], \tag{2.9}$$

$$k_F = (3\pi^2 n)^{1/3}, \tag{2.10}$$

with k_F being the Fermi wavelength. The GGA exchange energy is therefor given as

$$E_x^{\mathrm{GGA}}[n, \nabla n] = \int \varepsilon_x^{unif}(n) F_x(s) n(\boldsymbol{r}) d\boldsymbol{r}. \tag{2.11}$$

Similar to the LSDA the known properties of the uniform electron gas can be used as basis for constructing the enhancement factor.

[7] Further details on the different constraints that are fulfilled by the approximation can be found in Perdew and Schmidt [2001].

[8] See Perdew et al. [2005].

Specifically it is known that in the slowly varying density limit the expansion of $F_x(s)$ is given as

$$F_x(s) = 1 + \frac{10}{81}s^2 + \frac{146}{2025}\left(\frac{|\nabla^2 n|}{(2k_F)^2 n}\right)^2 + \cdots . \tag{2.12}$$

Using this expansion makes the GGA functional obey the LDA limit of $F_x(s = 0) = 1$. Secondly the constraint of the Lieb-Oxford lower bound can be used, which puts an upper limit for the exchange enhancement factor of $F_x(s) \quad \leq \quad 1.804$.[9]

The correlation depend on the LDA correlation similar to the exchange enhancement factor, but in a more complicated fashion.

The GGA approximation allows for a extrapolation away from the LDA limit to the slowly-varying high-density (small-s) regime. The GGA approximations have especially improved the approximations for molecular bonds, and only introducing a small extra computational cost to the calculations. There are however many limitations to the model still. The s-parameter provides a measure for the inhomogeneity, which the GGA functional approximations uses to assess how far the local density is from a uniform electron gas, and scale the uniform electron gas energy on basis of that. The electron density distribution of ionic and metallic solid crystals may however differ significantly,[10] and the gradient of the local density is not able to fully capture this. It is not possible by merely knowing the s value of a local density, if the electron is a part of a solid material or a molecular bond.

The DFA zoo is very populated with GGAs. Some functionals are found to have a better relative performance in describing solid state materials because they do not deviate too strongly from LSDA in the small s regime, while others have been optimized for theoretical chemistry while doing so deviate more from LSDA, thus making them less suitable for solid state materials studies. The most popular GGA functional is PBE and it's PW91 predecessor. A number of functionals are based variations of the PBE functionals, e.g.: PBEsol what uphold the slowly varying gradient expansion to full capacity of a GGA; and RPBE that has been targeted for chemisorption. The BLYP functional with it's combination of B88 exchange and LYP correlation was important early to show that GGA's could be used for chemistry. [11]

[9] See Lieb and Oxford [1981]

[10] See Csonka et al. [2009], Klimes et al. [2011]

[11] PBE and PW91 of Perdew et al. [1996a,b]; PBEsol and RPBE of Perdew et al. [2008], Hammer et al. [1999]; B88 and LYP of Becke [1988], Lee et al. [1988]

2.6 MGGA

At the third rung on Jacob's ladder the GGA is expanded by adding the Laplacian of the electron density and/or the kinetic energy density (KED) $\tau(\boldsymbol{r})$ of the occupied KS orbitals,

$$\tau_\sigma(\boldsymbol{r}) = \frac{1}{2}\sum_i^{\text{occ.}} |\nabla\phi_{i\sigma}(\boldsymbol{r})|^2.$$

The exchange of common MGGA functionals is similar to the GGA in the form of an enhancement factor to the LSDA energy, but including the before mentioned ingredients also. The KED can be expressed in a form as $\alpha = \left(\tau - \tau^{\mathrm{W}}\right)/\tau^{unif} \in [0, \infty]$, where $\tau^{\mathrm{W}} = |\nabla n|^2/8n$ is the von Weizsäcker KED, and $\tau^{unif} = \frac{3}{10}(3\pi^2)^{2/3}n^{5/3}$ is the KED of the uniform electron gas. The $\alpha = 0$ limit corresponds to $\tau = \tau^{\mathrm{W}}$, which is characteristic of electron densities with single-electron (iso-orbital) character, and for $\alpha = 1$ the KED is that of the uniform electron gas ($\tau = \tau^{unif}$). The local KED is therefore able to discriminate between these two very different regimens in terms of what kind of bonding is taking place. That the MGGA in such a direct way is able to classify densities, has been given as cause for it's better performance than GGA. With the KED it is possible to fulfill the slowly varying density expansion to a higher order than for GGA, due to the extra flexibility of the functionals, which provides means to restrict the extended functional form.

The MGGAs are considered semi-local functionals as they rely on the occupied KS orbitals which are readily available in DFT calculations. The extra computational overhead of calculating the total energy of a MGGA functional is modest compared to a GGA functional, when using the Neumann et al. [1996] method.

A number of MGGA functionals have been proposed but many fewer than GGAs. Of common functionals are VSXC, TPSS and it's revTPSS revision, and M06-L. Lately the MS-0, MS-1 and MS-2 have been proposed which insure that more of the exact constraints are fulfilled.[12]

2.7 The non-local functionals for dispersion

The dispersion force is an inherent long-range effect, and it can exist between fragments that do not have any density-density overlap. The local to semi-local functional type listed above (LDA, GGA, and MGGA) should therefore by construction not be able to capture dispersion effects, and this is also observed in practice.[13]

[12] VSXC is of Van Voorhis and Scuseria [1998]; TPSS and revTPSS are of Tao et al. [2003], Perdew et al. [2009]; M06-L is of Zhao and Truhlar [2006]; MS-0 is from Sun et al. [2012a], while MS-1 and MS-2 are of Sun et al. [2013].

[13] See Kristyán and Pulay [1994]

In the following will be presented the main ideas of the Rutgers-Chalmers (RC) non-local correlation approximation, and it's variations. Following by examples of semi-local density functionals where this correlation functional has been introduced to offer applicable functionals for describing vdW systems.

A number of non-density-density based vdW correction methods have also been suggested, where the dispersion is based on summing up approximations for pairwise dispersion interaction between atoms. The dispersion energy is then added to the kohn-sham energy post the solution of the Kohn-Sham equations. These methods have successfully been employed in many molecular system calculations, but they are however not in general suitable for solids.[14]

Some notable examples of these are the DFT-D method, the TS09

2.7.1 *Rutgers-Chalmers non-local correlation approximation*

With the Rutgers-Chalmers (RC) correlation approximation a correction to the LDA correlation energy for non-uniform electron densities is introduced. The RC correlation approximation is therefore not only a vdW approximation, but serves the role of covering all non-local type correlation. Other variants of the vdW density functionals have been made to be combined with semi-local correlation functionals, which will later be discussed.[15]

The starting point is the adiabatic connection fluctuation-dissipation (ACFD) formalism to the exact ground state correlation energy. From ACFD expression an approximation to the non-local part of the correlation can be made in the compact form

$$E_c^{\mathrm{nl}} = \frac{1}{2} \iint n(\boldsymbol{r})\phi(\boldsymbol{r}, n(\boldsymbol{r}), \nabla n; \boldsymbol{r}', n(\boldsymbol{r}'), \nabla n(\boldsymbol{r})) n(\boldsymbol{r}') d\boldsymbol{r} d\boldsymbol{r}', \qquad (2.13)$$

which is a 6-dimensional integral over the interaction kernel ϕ, and so depends on the densities and density gradients in all pairs of spacial points $\boldsymbol{r}$ and $\boldsymbol{r}'$.[16] The resulting interaction kernel, ϕ, has several appealing features. Since the asymptotic form is

$$\phi \sim |\boldsymbol{r} - \boldsymbol{r}'|^{-6} \text{ for } |\boldsymbol{r} - \boldsymbol{r}'| \to \infty, \qquad (2.14)$$

the nonlocal correlation will follow the $-C_6 R^{-6}$ formulation for well-separated densities, that is missing in the local approximations. The kernel's symmetric properties furthermore results in $E_c^{nl} = 0$ for $\nabla n(\boldsymbol{r}) = 0$, such that the non-local correlation vanishes for the uniform electron gas. The kernel includes a local wave vector, usually

[14] See Tkatchenko and Scheffler [2009], Grimme et al. [2010], Sato and Nakai [2009], Becke and Johnson [2005, 2006].

[15] Only a very broad overview to the method is given here. For more detail consult Dobson and Gould [2012].

[16] A number of approximation have been introduced to bring the ACFD to the form of 2.13, which are omitted here as the level of detail needed to justify the approach is beyond the scope of this chapter.

denoted $q_0(\boldsymbol{r})$, for which a density gradient dependence enters. It is defined as a modulation of the fermi wave vector through

$$q_0(\boldsymbol{r}) = \frac{\epsilon_{ex}^0}{\epsilon_x^{LDA}} k_F(\boldsymbol{r}), \tag{2.15}$$

where $\epsilon_x^{LDA} = \epsilon_x^{LDA}[n]$ and $\epsilon_{xc}^0 = \epsilon_x^0[n, \nabla n]$ is given by a a gradient correction to LDA correlation of by:

$$\epsilon_{xc}^0 = \epsilon_{xc}^{LDA} - \epsilon_x^{LDA} \left[\frac{Z}{9} \left(\frac{\nabla n}{k_F n} \right)^2 \right], \tag{2.16}$$

with $Z = -0.8491$ in the original form. The non-local correlation also contributes to the exchange-correlation potential through $v_c^{nl} = \delta E_c^{nl} / \delta n$, such that fully self-consistent calculations are possible.

In the form outlines above, a 6D spacial integral of all the density pairs need to be made, which is very computational expensive as the system at hand grows. It has been found that the kernel form can be cast in a slightly different form, that allows for a fast Fourier transformed procedure. This procedure by Román-Pérez and Soler [2009] allows for an implementation of the functional that is significantly faster, such that the non-local correlation functional becomes computational feasible for most small relevant systems.

The RC none-local correlation was paired with the revPBE exchange, chosen to mimic exact exchange, and the LDA local correlation in vdW-DF exchange correlation functional. It was found that the non-local correlation cannot be paired successfully with the exact exchange though, and the error cancelation effect of the RC non-local approximation with exchange is thus found needed to yield good results. [17]

To improve on the results of the vdW-DF functionals it has been suggested to pair the non-local correlation with other exchange functionals. These functional types includes the optPBE-vdW, optB88-vdW and optB86b, where the first two have been optimized through the use of the S22 benchmark dataset for non-covalent interactions, and the latter for bulk energies. Another vdW type functional is that of C09-vdW that in the exchange matches a enhancement factor form in the low reduced gradient (s) that fulfills the slowly varying electron gas, with revPBE exchange enhancement factor for high s values.[18]

Later the vdW-DF functional have been revised to the vdW-DF2 functional, see Lee et al. [2010], by changing the exchange functional that the non-local correlation is combined with, and by employing a larger gradient dependence in q_0 through choosing $Z = -1.887$ in

[17] See Dion et al. [2004], Langreth et al. [2009]

[18] For optPBE-vdW, optB88-vdW and optB86b-vdW see Klimes et al. [2010, 2011]; for C09-vdW see Cooper [2010].

equation 2.16.

2.7.2 VV09 and VV10 non-local correlation approximations

Another approach to approximating the ACFD equation for the non-local correlation functional has been done by Vydrov and Van Voorhis in the studies towards the VV09 and VV10 non-local correlation functionals.[19] The VV10, build on ideas from VV09, has been found to yield very good performance on non-covalently and covalently bonded molecular systems. In Sabatini et al. [2013] the VV10 kernel was furthermore reformulation so that a Fourier transformation could be made similar to the original non-local functional, which makes the revised VV10 (rVV10) computational feasible for relevant system sizes.

2.8 Hyper-GGA and hybrids

The forth-rung density functionals introduce the non-local occupied KS orbitals. With the occupied non-local KS orbitals it is possible to calculate the exact exchange (EXX) functional. The formal definition of the correlation is therefore what is left to describe the exact density functional after the exact exchange has been subtracted. With the exact exchange given, only an approximation to the exact correlation functional is needed, but to find a suitable correlation functional has however not been fruitful, and different approaches has therefore been made where only part of the exact exchange is used.

The hybrid functionals mixes a fraction of the exact exchange energy with that of lower-rung DFAs, and with this it is possible to achieve much better energetics. This shows how error cancelation between the approximations to the exchange and correlation is important for the performance of the lower-rung functionals. The hybrid functionals are very popular for quantum chemistry because of their good description of molecular thermochemistry.[20]

The long-ranged Coulomb potential however renders hybrid DFAs very computational demanding for periodic systems, especially metallic systems. Screening of either the long-range og the short-range part of the Coulomb potential for EXX can be used to improve the characteristics of the hybrids; either performance wise or make them less computational demanding. On how the range-separation can improve performance will be given in the the below section on self-interaction error.

The most popular functional of the hybrid type is that of B3LYP.

[19] See Vydrov and Van Voorhis [2009, 2010a], Langreth and Lundqvist [2010], Vydrov and Van Voorhis [2010b,?]

[20] See Becke [1993, 1997], Zhao and Truhlar [2008].

Of others it is worth mentioning PBE0 and HSE.[21]

2.9 Random Phase approximation + corrections

At the highest level of Jacobs ladder is the random phase approximation (RPA), which is based on the the second order perturbation theory. The RPA is therefore a natural companion to the exact exchange. A description of the van der Waals interaction naturally comes out of the formalism, which makes the approximation useful for benchmarking vdW functionals. The approximation has been proven to give qualitative better descriptions than semi-local functional for systems where the lower rung functionals fails to give the correct qualitative description, e.g. in what site CO will adsorb on transition metal surfaces.[22] It has however been observed that short range correlation is performing poorly, and screening is therefore suggested. The calculations are much more computational expensive than for the semi-local functionals but implementations optimized for GPU's significantly lower the computational costs of the method, and bring it within reach for many current studies. The RPA is seldom self-consistent but added on top of orbitals from semi-local functional calculations.[23]

With the rALDA reformulation of the RPA method, RPA now also exceeds the performance of standard semi-local functionals for most types of energetics.[24]

A role for the RPA method with it's high accuracy, but also high computational cost, is as a verification of lower rung functionals and to benchmark calculations where either high quality quantum chemistry methods are not feasible and where high quality experimental data are not available.

2.10 The self-interaction error

In the Kohn-Sham equations for the energy (equation 2.5), it was observed in the early days of quantum mechanics that the Hartree energy, given as

$$J[n] = \frac{1}{2}\int d\boldsymbol{r}\int d\boldsymbol{r}'\frac{n(\boldsymbol{r})n(\boldsymbol{r}')}{|\boldsymbol{r}'-\boldsymbol{r}|} \tag{2.17}$$

did not vanish for one-electron system, due to a spurious self-interaction (SI) inherent in it. For Hartree-Fock theory, i.e. exact exchange of rung 4 and up, this is not a problem as the self-interaction term is cancelled by the exchange self-interaction terms. This is however not so for the semi-local exchange functionals (LDA, GGA and MGGA), and this so called self-interaction error (SIE) is believed to

[21] B3LYP is in Stephens et al. [1994]; PBE0 is in Adamo and Barone [1999]; and HSE is in Heyd et al. [2003].

[22] See Schimka et al. [2010], Olsen et al. [2011]

[23] See Yan et al. [2013]

[24] See Olsen and Thygesen [2013]

be the cause of many of the failures of approximate density functionals.[25] The goal has therefore been to find methods for these lower rung functionals that effectively removes the SIE, both for chemistry and solids, while still make these functionals computational less demanding than the 4th and 5th rung functionals. The SIE has been shown to be the reason for different failings of low-rung functionals, these include: Not producing derivative discontinuities at integer electron numbers of the total energy as a function of the orbital filling; also the failure to reproduce localized orbitals; and to provide proper band-gabs for many materials. These failures are as mentioned all different reminecense of the fundamental SIE.

A number of self-interaction correction (SIC) schemes have been suggested. Most notable is the PZ-SIC of Perdew and Zunger [1981], that formally is correct. However the method fails in many cases, and it is not clear how this formalism can be made widely used. [26] A number of new methods methods tries to alleviate these problems with the PZ-SIC, but non have been widely adopted.[27]

The Hubbard U correction within the linear response methods, provides another way of removing the self-interaction for lower rung functionals.[28] This method ensures that the Hubbard U term is not a fitting parameter, but chosen instead to minimize self-interaction errors. A number of extension have been proposed for also correcting for the inter-site non-linearity of the energy when filling up orbitals, and further to calculate forces so that different energies can be compared.[29]

For the hybrid functionals where only partial EXX is used a SIE will be present. For these functionals long-range-corrected form has been proposed, where the long-range part of the exchange is used to remove the self-interaction, while the short range exchange is that of a lower rung approximation. The short range semi-local exchange insures that proper error cancelation with the lower-rung correlation functional can happen.[30] The long-range part of the EXX is unfortunately, as previous mentioned, the most computational expensive part of EXX.

Another approach is to create an model potential that will provide derivative discontinuity or other desired properties. Much success have been obtained with such an approach to calculate bandgabs in materials; using for instance the GLLB potential; however this potential, and many others, fail to be integrable to a energy functional, and can thus only be used as a post SCF approach.[31] In Armiento and

[25] See Vydrov and Scuseria [2004]

[26] Pederson and Perdew [2011]

[27] See for instance Dabo et al. [2010]

[28] Cococcioni and de Gironcoli [2005]

[29] See respectively:Campo and Cococcioni [2010], Kulik and Marzari [2011]

[30] See Krukau et al. [2008]

[31] For GLLB see Krukau et al. [2008]. For discussion on different approaches to create energy functionals from model potentials see Elkind and Staroverov [2012].

Kümmel [2013] is has however been shown it is possible to create a GGA exchange energy functional with derivative discontinuity. This functional has a form that is very different from usual GGA functionals and it is to be seen how to create a functional with the necessary features to reproduce derivative discontinuities, and at the same time provides good energetics.

2.11 *Solving the Kohn-Sham equations computationally efficiently*

The following will be an account for the calculations performed towards creating the data given in the thesis.

Given a exchange-correlation energy functional, the ground state density and total energy are calculated using the iterative self-consistent field procedure (SCF): Initial a starting density is used to calculate the KS effective potential $v_s(\boldsymbol{r})$ for all electrons, and the solution of the Kohn-Sham eigenvalue problem (equation 2.2). The single particle solutions $\psi'_{\alpha\sigma}$ will however correspond to a different density $n'(\boldsymbol{r}) = \sum_{\alpha,\sigma} |\psi'_{\alpha\sigma}|^2$ than the initial; this output density is now used for solving the eigenvalue problem again, and repeat the process. The SCF loop continues until convergence is reached, by which the densities and total energy do not change significantly between iterations.

To solve the Kohn-Sham equations, a representation of the electronic densities, potentials and wavefunctions are needed. This representation should provide adequate description, while being efficient in terms of storage and computational cost. Common basis-sets for the representation are atomic-centered orbitals, plane waves, and real-space grid. The atomic wavefunctions are eigenfunctions to the atomic Hamilton, so they are to be mutually orthogonal. The atomic core states are localized to the core, and are very different from the delocalized valence orbitals. For the valence electrons to be orthogonal to the core orbitals, they therefore have to be rapidly oscillatory in the core region, and that makes them expensive to represent computationally. Different approaches are used to go about this.[32]

The core orbitals change very little with the chemical environment for most systems, and a simple approximation is therefore the frozen core approximation, where the core orbitals are not relaxed in the SCF procedure. The frozen-core approximation is usually implemented through pseudopotentials, that are smoothly varying potentials constructed to mimic the effect of the ion and the core electrons on the valence electrons. Using pseudopotentials one only solves the Kohn-Sham equations for the valence electrons. This method is computational much cheaper than solving for the core electrons, but one

[32] See Martin [2004]

also discard all information about the Kohn-Sham orbitals close to the core.

A way to go about discarding all core orbital information, is by using an all-electron method such as the augmented plane wave (APW) method. The APW method divides the space up in a regions for the atomic core and for the interstitial regions. The core is then treated with atom-centered augmentation spheres in which the wave functions are taken as atomic-like partial waves, to efficiently reproduce the rapid oscillations. The interstitial regions are instead expanded with smoothly varying envelope functions, e.g., plane waves. The partial waves and the envelope functions are then matches at the augmentation sphere boundaries.[33]

The Projector Augmented Wave formalism (PAW) of Blöchl [1994] extend on the augmented-wave methods and the pseudopotential approach.[34] In the PAW method the rapidly oscillating wave functions of the core are linear transformed into auxiliary wavefunctions, and atomic corrections are then added inside the augmentation spheres. The Kohn-Sham equation can then be solved independently in the two regions, and the solution of the smooth part and the corrections for the atomic are then added together afterwards for the potentials and the densities to give the all-electron solution.

The DFT calculations presented and used later as inputs for the fitting routines, were calculated using the GPAW software package, which employs the PAW method, within a real grid or plane waves representation. The Atomic Simulation Environment was furthermore used as interface to GPAW.[35]

[33] See Slater [1937], Martin [2004], Blöchl [1994]

[34] See Kresse and Joubert [1999]

[35] See Enkovaara et al. [2010], Bahn and Jacobsen [2002]

3 Datasets for training and validation

For the training of the exchange correlation model and following validation it is important with reliable compilations of materials properties and chemical observables. These datasets should represent the condensed matter interactions that is DFT studies, and the reference data should be of the highest quality. This chapter presents datasets to be used throughout the rest of this thesis.

3.1 Datasets of materials properties

The benchmark data can be from either experimental studies or calculated from high-level theory such as CCSD(T).[1] The most elaborate wavefunction methods are very accurate, and can be considered essentially exact for molecular properties, and they are thus very good for benchmarking. The CCSD(T) "model chemistry" has become the standard for benchmark data to be directly compared to DFT results. The experimental data are often not as accurate as theoretical benchmarks, but many atomistic systems are simply impossible to treat with expensive wavefunction theory, e.g., the extended lattices of solid crystals. To capture materials properties for these systems one must resort to careful experiments, preferably at low temperature and possibly with extrapolation to the zero Kelvin limit and corrections for zero-point motion of atoms.

The following benchmark datasets are either adapted from literature or compiled from published works.

3.1.1 G3/99 and G2/97: Molecular formation energies

The molecular formation enthalpies of the G3/99 thermochemical test set of Curtiss et al. [1997] represent intramolecular bond energetics. It has become very popular for benchmarking and calibrating electronic structure methods. The 223 molecules may be divided into three subsets denoted G3-1, G3-2, and G3-3 comprising 55, 93, and 75 molecules, respectively. The G3-1 and G3-2 subsets constitute

[1] The CCSD(T) method is a quantum chemistry method, where the many-body Schrödinger equation is solved very accurately.

G2/97, in which case the two subsets may be denoted G2-1 and G2-2, respectively.

The formation enthalpies are experimentally determined. In accordance with the procedure of Curtiss et al. [1997] they are extrapolated to zero Kelvin by correcting for thermal and vibrational contributions. Thermal corrections and zero-point energies from Curtiss et al. [1997] and Staroverov et al. [2003] are used. The result is 233 electronic-only static-nuclei formation energies $\Delta_f E$, i.e., negatively signed atomization energies, which are directly comparable to predictions from ground state DFT. Contributions to $\Delta_f E$ from spin-orbit effects are not corrected for. This is expected to be of little overall consequence.[2]

Theoretical G3/99 formation energies are calculated from the difference between molecular and atomic total energies as

$$\Delta_f E = E_M - \sum_A E_A, \tag{3.1}$$

where A runs over all atoms in the molecule M, while E_M and E_A are ground state molecular and atomic total energies, respectively.

3.2 RE42: Molecular reaction energies

The RE42 compilation contains 42 zero-Kelvin reaction energies involving 45 different molecules from G3/99, and was presented in Wellendorff et al. [2012]. The theoretical reaction energies are calculated from total electronic energies as $\Delta_r E = \sum_P E_P - \sum_R E_R$, where the sums run over reactant (R) and product (P) molecules.

3.3 DBH24/08: Molecular reaction barriers

The chemical reactant and product states are often separated by an energy barrier, which must be surmounted if the reaction is to proceed. The DBH24/08 set of Zheng et al. [2009] comprises 12 forward (V_f) and 12 backward (V_b) benchmark barriers

Ground- and transition-state molecular geometries, are calculated using the quadratic configuration interaction with single and double excitations (QCISD) wavefunction method, are from Zheng et al. [2007]. Density functional barrier heights are computed from the transition state total energy ($E_‡$) and the initial (E_i) and final (E_f) state total energies as

$$\begin{aligned} V_f &= E_‡ - E_i, & (3.2) \\ V_b &= E_‡ - E_f. & (3.3) \end{aligned}$$

[2] See Curtiss et al. [1997]

3.4 S22 and S22x5: Non-covalent bonding

The S22 dataset of Jurecka et al. [2006] represents van der Waals interactions and hydrogen bonding by considering non-covalent bonding between molecular dimers and complexes. It has been widely used for assessment[3] and parametrization[4] of density functional methods for vdW type interactions. The datasets has however now been somewhat superseded by the newer and larger S66 set of Rezac et al. [2011]. The S22 set consists of CCSD(T) interaction energies between relatively small molecular complexes, but includes also non-covalent bonding between the somewhat larger DNA and RNA bases adenine, thymine, and uracil, as well as 2-pyridoxine and 2-aminopyridine. The 22 complexes are divided into three groups according to the type of interaction predominantly responsible for stabilizing the complex; hydrogen bonding, dispersion interactions, and a mixture of dispersion and electrostatic interactions. This categorization was made on the basis of interaction energy decompositions using the symmetry-adapted perturbation theory method.

MP2 or CCSD(T) geometries at equilibrium intermolecular separations from the original work in Jurecka et al. [2006]are used. Benchmark CCSD(T) interaction energies with extrapolation to the complete basis set (CBS) limit were reported in that same publication. However, most likely due to the computing resources available at the time, different basis sets were used for small and large complexes. Later works have therefore revised the S22 interaction energies, employing larger and identical basis sets for all complexes at the original geometries.[5] For the larger complexes the reported basis set effects are significant, so the CCSD(T)/CBS energies of Takatani et al. [2010] are adopted here as the current best-estimate of the true S22 interaction energies.

An extension in form of the S22x5 extension was proposed in Grafova et al. [2010]. In addition to the near-equilibrium intermolecular distances, S22x5 contains for each complex four non-equilibrium binding energies. Thus, CCSD(T) potential-energy curves (PECs) for each complex are mapped out at relative interaction distances d of 0.9, 1.0, 1.2, 1.5, and 2.0 as compared to S22. We here divide S22x5 into five subsets according to interaction distance, e.g., "S22x5-0.9".

The computational procedure used for S22x5 was identical to the S22 one, so we expect the basis set deficiencies to persist in S22x5. The non-equilibrium data points on each PEC are therefore corrected according to the difference between original and revised S22x5-

[3] See: Gulans et al. [2009], Cooper [2010], Kannemann and Becke [2010], Sherrill [2010], Hanke [2011], Goerigk and Grimme [2011], Zhao and Truhlar [2008].

[4] See Kannemann and Becke [2010], Klimes et al. [2010], Lee et al. [2010], Vydrov and Van Voorhis [2010], and Grimme et al. [2010], Zhao and Truhlar [2006].

[5] See Takatani et al. [2010] and Podeszwa et al. [2010].

1.0 CCSD(T) energies. The proposed correction were published in Wellendorff et al. [2012]. These are very small on average but significant for certain larger complexes. The modified CCSD(T) interaction energies are used throughout for the S22x5 dataset and subsets.

Each S22x5 density functional interaction energy E^d_{int} is computed as the difference between the total electronic energy of the interacting complex E^d_0 and those of its two isolated molecular constituents, E^d_1 and E^d_2,

$$E^d_{\text{int}} = E^d_0 - E^d_1 - E^d_2. \tag{3.4}$$

Computational accuracy is enhanced by keeping all atoms in the monomers in the same positions in the supercell as those atoms have when evaluating the total energy of the complex. With the sign convention in Grafova et al. [2010] stable intermolecular bonding is here taken to mean negative interaction energy.

3.5 Crystalline solids

We represent the energetic and structural properties of crystalline solids in the follow datasets of experimental data.

3.5.1 Sol34Ec

Cohesive energies of 34 Period 2–6 pure crystals in fcc, bcc, diamond, and hcp lattices. Zero-point effects are not considered. This dataset was used in Wellendorff et al. [2012], where the included systems are listed.

3.5.2 Sol27

It was shown by Csonka et al. [2009] that removal of thermal and zero-point contributions to experimentally determined lattice constants and bulk moduli may be important when benchmarking density functional methods. Experimental zero-Kelvin lattice constants and cohesive energies (E_c) contain zero-point vibrational contributions, leading to zero-point anharmonic expansion (ZPAE) of the lattice and zero-point vibrational energy (ZPVE) contributions to E_c. As discussed in Alchagirov et al. [2001], an estimate of the ZPVE may be obtained from the Debye temperature Θ_{D} of the solid according to

$$\text{ZPVE} = -\frac{9}{8} k_{\text{B}} \Theta_{\text{D}}. \tag{3.5}$$

The vibrational contribution is subtracted from the cohesive energy, leading to increased stability of the crystal towards atomization.

The same reference derived a semi-empirical estimate of the ZPAE contribution to the volume of cubic crystals.

The Sol27LC and Sol27Ec sets of zero Kelvin lattice constants and cohesive energies of 27 fcc, bcc, and diamond structured bulk solids are appropriately corrected for zero-point phonon effects. These datasets were also used in Wellendorff et al. [2012].

3.5.3 *Extended solids dataset of Sol54Ec and Sol58LC*

The Sol58 lattice constants (Sol58LC) and Sol54 cohesive energies (Sol54Ec), are extensions of the Sol27 sets to include also mixed-element compounds in the rock-salt, cesium chloride, and zincblende cubic crystal structures. The low-temperature zero-point exclusive data are from Schimka et al. [2011], Haas et al. [2009] (Sol58LC) and Schimka et al. [2011] (Sol54Ec), respectively.

The crystal cohesive energy for a given lattice constant a is calculated from

$$E_c = E_A - E_B, \tag{3.6}$$

where E_A is the total energy of the free atom and E_B the bulk total energy per atom. The equilibrium (maximum) cohesive energy of a stable solid is thus a positive quantity. Equilibrium lattice constants a_0 are determined from fitting the SJEOS equation of state to cohesive energies sampled in five points in a small interval around the maximum of the $E_c(a)$ curve, see Alchagirov et al. [2001]. For hcp-structured crystals the c/a lattice constant ratio is fixed at the experimental one.

3.6 *CE27: Chemisorption on solid surfaces*

The CE27 datasets contains chemisorption energies of simple molecules on late transition-metal surfaces. They are derived from temperature programmed desorption experiments or from microcalorimetry, most often at low coverage. The 27 chemisorption energies have been critically chosen from literature with emphasis on reliability as well as covering a reasonably wide range of substrates and adsorbates. CE17 is a subset of CE27. Details regarding adsorption mode, adsorption site, references and computational setups can be found in Wellendorff et al. [2012].

Most of the surface reactions are associative adsorption processes at 0.25~ML coverage. In that case the chemisorption energy ΔE is computed according to

$$\Delta E = E_{AM} - E_M - xE_A, \tag{3.7}$$

where E_{AM} is the total electronic energy of the adsorbate A on metal surface M, and E_A and E_M total energies of the isolated adsorbate and metal surface, respectively. The constant x equals 1 for associative adsorption and N_2 dissociation on Fe(100), while $x = \frac{1}{2}$ for dissociative H_2 chemisorption. In the case of NO dissociation on Ni(100) at 0.25~ML coverage the chemisorption energy is

$$\Delta E = E_{AM} + E_{BM} - 2E_M - E_{AB}, \tag{3.8}$$

where AB is the NO molecule.

4 Machine learning methods

In this chapter we will introduce the different machine learning(ML) tools used for fitting the Bayesian Error Estimation functionals, which for brevity are called BEE functionals or for short BEEFs. These ML tools have been introduced at different points in the history of the BEEF family development and this chapter is intended to create a more coherent introduction to the ML tools used. The goal is however not to provide an in depth statistical foundation, but rather to present the methods concisely so to create a good overview.

4.1 Parametrization

To fit a exchange-correlation functional, one must parametrize the functional space. For the exchange-enhancement factor this is usually achieved through a parametrization in the normalized gradient density s within the GGA formalism. A parametrization is sought that can describe the optimal model with the fewest number of parameters.

The general formulation of a linear parametrization model is

$$\mathcal{M}(x, \boldsymbol{a}) = \boldsymbol{a}_0 f_0(x) + \boldsymbol{a}_1 f_1(x) + ... + \boldsymbol{a}_p f_p(x) = \sum_{p=0}^{N_p} \boldsymbol{a}_p f_p(x), \qquad (4.1)$$

where x is the input, $\boldsymbol{a}$ is the coefficients to the model and $f(x)$ are the basis functions. For the simple polynomial series the basis functions are given as $f_p = x^p$.

The polynomial series is however not orthogonal, and we will therefore later use the Legendre polynomial series, which is orthonormal over the region $[-1, 1]$. We can use a Padé approximant to transform a dimension from $[0, \infty[$ to $[-1, 1]$.

4.2 The Bayesian connection

4.2.1 Probability theory and Bayes' theorem

The foundation of the Bayesian inference is through Bayes' theorem that provides a connection between the probability distribution of A and B, given as $P(A)$ and $P(B)$, and the conditional probabilities of A given B, or B given A, hence $P(A|B)$ and $P(B|A)$ respectively.

First we define the product rule, which says that the conditional probability $P(A|B)$ is given by the probability joint probability of A and B by

$$P(A \cap B) = P(A, B) = P(A|B)/P(B). \tag{4.2}$$

Secondly, the sum rule enable us to find the probability of A by summing up all the joint probabilities.

$$P(A) = \sum_B P(A, B), \tag{4.3}$$

where the sum runs over all probability distributions of B that link to A. Now using these two rules enable use to define Bayes's theorem in the common form

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)}, \tag{4.4}$$

so that A can be deduced by B, using the prior knowledge of the distributions of A and B given as $P(A)$ and $P(B)$.

Lets consider the case where a dataset $\mathcal{D}$, which we will attempt to describe by the model $\mathcal{M}$ with the parameter vector $\boldsymbol{a}$. It is assumed that $\mathcal{D}$ has been generated by $\mathcal{M}$, but that $\boldsymbol{a}$ is unknown. And that the data has an added layer of noise ($\mathcal{D} = \mathcal{M} + \epsilon$). Our goal is now to find **a** given $\mathcal{D}$ and our knowledge about the noise ϵ. Using Bayes's theorem[1] we have that the posterior probability $P(\boldsymbol{a}M|\mathcal{D})$ is the likelihood $\frac{P(\mathcal{D}|\boldsymbol{a}M)}{P(\mathcal{D})}$ times the prior distribution $P(\boldsymbol{a})$, hence

$$P(\boldsymbol{a}M|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{a}M)P(\boldsymbol{a})}{P(\mathcal{D})}. \tag{4.5}$$

The denominator in the above expression can be described as

$$P(\mathcal{D}) = \int P(\mathcal{D}|\boldsymbol{a})P(\boldsymbol{a})d\boldsymbol{a}, \tag{4.6}$$

and is a normalization factor that insures that the posterior distribution integrates to one. In the following example we will find the best model to describe the given data.

[1] posterior $\propto$ likelihood $\times$ prior

4.2.2 *Maximum-likelihood model for model with Gaussian noise*[2]

Assuming that the model has an overlay of Gaussian noise, then the likelihood has the form

$$P(\mathcal{D}|\mathcal{M}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n|\mathcal{M}(\boldsymbol{x}_n, \boldsymbol{a}), \boldsymbol{\beta}^{-1}\right), \tag{4.7}$$

where $\boldsymbol{\beta}$ is the precision matrix, x_n is measurement variables to the model and y_n is the associated target vector generated from the model. The precision matrix $\boldsymbol{\beta}$ is given as the variance for the noise process. $\mathcal{N}(y_n|\mu, \sigma^2)$ would denote a Gaussian distribution on top of y_m with a mean μ and variance σ^2, hence $\boldsymbol{\beta} = \beta\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

Assuming that we have an equal prior expectation to any $\boldsymbol{a}$, then we can maximize the likelihood of the model, which is the same as maximizing the logarithm to of the likelihood function, hence

$$\underset{\boldsymbol{a}}{\operatorname{argmax}}\, P(\boldsymbol{a}|\mathcal{D}) = \underset{\boldsymbol{a}}{\operatorname{argmax}}\, \ln P(\boldsymbol{a}|\mathcal{D}), \tag{4.8}$$

$$\ln P(\boldsymbol{a}|\mathcal{D}) = -\frac{\beta}{2}\sum_{i}^{N_d}\{y_i - \mathcal{M}(\boldsymbol{a}, x_i)\}^2 \tag{4.9}$$

which is similar to minimizing the conventional least squares(LS) loss function.

$$L(\boldsymbol{a}) = \frac{1}{2}\sum_{i}^{N_d}(\mathcal{M}(\boldsymbol{a}, x_i) - y_i)^2. \tag{4.10}$$

The solution of the LS for the above problem is also known from Gauss-Markov theorem.[3] And if the precision matrix $\boldsymbol{\beta} \neq \beta\boldsymbol{I}$, then the maximum likelihood solution correspond to that of the generalized least squares solution, which also fulfills the Gauss Markov theorem.

[2] See [Bishop, 2006]

[3] See p.51 in [Hastie et al., 2009] for definition of Gauss-Markov theorem.

4.2.3 *Overfitting*

Now for finding the underlying model of the noisy data of last section, the number of parameters N_p in the underlying model could be unknown. It is illustrated in figure 4.1 below what can happens when we increase the number of parameters for the model. All the data points will be very well described but the fitted model does not describe the underlying model. This is called an overfit.

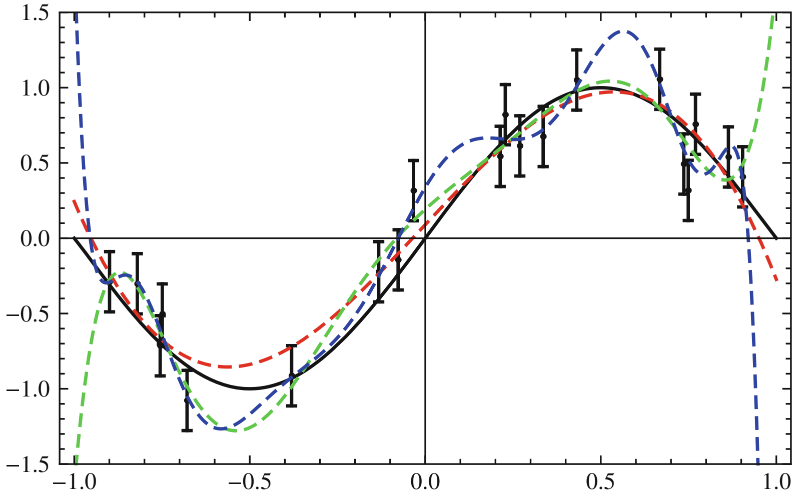


Figure 4.1: The black curve shows a sine function, from which a number of measurements with associated error are shown by the black error bars. The red, green, and blue curves show the optimal polynomial fits of degrees 3, 7, 11 respectively. The deviation of the higher order fits are are a clear sign of overfitting. Adopted from [Petzold et al., 2012].

4.2.4 *Prior model expectation and regularization*

So our goal is to find the underlying model, but we did not use our prior knowledge fully, as we had an expectation for that the model should not be too complex. We now assume a Gaussian prior distribution for the model parameters given as

$$P(\boldsymbol{a}|M\omega) = \mathcal{N}(\boldsymbol{a}|\boldsymbol{0}, \omega^{-1}\boldsymbol{I}) = \left(\frac{\omega^2}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\omega^2}{2}\boldsymbol{a}^T\boldsymbol{a}\right\}, \quad (4.11)$$

where ω is the precision of the model distribution. This posterior describes an uncertainty distribution the parameters in the model, meaning that if we did not have any data at hand then we would assume a zero vector solution ($\boldsymbol{a} = \boldsymbol{0}$). With data at hand our prior expectation will then compete with likelihood probability. The logarithm to the posterior distribution now gives

$$\ln P(\boldsymbol{a}|\mathcal{D}) = -\frac{\beta}{2}\sum_{i}^{N_d}\{y_i - \mathcal{M}(\boldsymbol{a}, x_i)\}^2 - \frac{\omega^2}{2}\boldsymbol{a}^T\boldsymbol{a} + const, \quad (4.12)$$

which we maximize to find the optimal model parameters $\boldsymbol{a}$.

4.2.5 The frequentist interpretation

The maximum likelihood problem of equation 4.12 above can be translated to language of the frequentist by introducing the cost function as

$$C = -\ln P(\boldsymbol{a}|\mathcal{D}) = L(\boldsymbol{a}, \mathcal{D}) + R(\boldsymbol{a}, \omega) + const, \tag{4.13}$$

which we wish to minimize. L is as mentioned earlier the loss and R is the regularization. The regularization term penalizes large coefficients, and the loss term penalizing not fitting the data well. The regularization term found in equation 4.12, $R = \frac{1}{2}\omega^2 \boldsymbol{a}^T \boldsymbol{a}$, is generally known by the names ridge regression, shrinkage and weight decay. ω is the regularization strength. We will abbreviate the least squares loss function with ridge regression regularization as RR-LS.

A more general regularizer can be used in the form of

$$R = \frac{\omega^2}{2} \sum_p^{N_p} |a_p|^q, \tag{4.14}$$

where $q = 2$ is that of the quadratic regularizer introduced above as ridge regression. $q = 1$ is know as the lasso, and this penalization has the effect that for a large regularization strength some of the coefficients will be driven to zero, and thus making the model more sparse. In the following we will limit ourselves to $q = 2$, but later we will change the norm of the loss function.

4.3 Minimizing the cost function

We now want to minimize the least squares ridge regression (RR-LS) cost function given as

$$C(\boldsymbol{a}) = \frac{1}{2}(\boldsymbol{X}\boldsymbol{a} - \boldsymbol{y})^2 + \frac{1}{2}\omega^2 \boldsymbol{a}^T \boldsymbol{a}, \tag{4.15}$$

where we have introduced the design matrix $\boldsymbol{X}$, an $N \times M$ matrix, with row i given as the parametrization series $f_p(x)$ of data point i. We can write the loss in this manner because the loss is linear in the coefficient vector $\boldsymbol{a}$. It is noted that RR-LS cost is quadratic in $\boldsymbol{a}$, and a close form solution will therefore exist.

To minimize the cost function we look for where the gradient in the solution space is zero, hence:

$$\begin{aligned}
\frac{\partial C}{\partial \boldsymbol{a}} &= 0 \Rightarrow && (4.16)\\
0 &= \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{a}_0 - \boldsymbol{X}^T y + \omega^2 \mathbf{a}_0 \Leftrightarrow && (4.17)\\
\boldsymbol{a}_0 &= \left(\boldsymbol{X}^T \boldsymbol{X} + \omega^2 \boldsymbol{I} \right)^{-1} \boldsymbol{X}^T y, && (4.18)
\end{aligned}$$

where the solution vector $\boldsymbol{a}_0$ is denote the coefficient vector $\boldsymbol{a}$ that minimizes the cost. The singular value decomposition (SVD) of $\boldsymbol{X}$ is given as $\boldsymbol{U\Sigma V}^T$, where $\boldsymbol{\Sigma}$ is the singular values matrix with the singular values Σ_p in the diagonal, and $\boldsymbol{V}^T$ is the right singular vectors matrix. $\boldsymbol{V}$ is a unitary transformation, hence $\boldsymbol{V}^T \boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. We can use the SVD to rewrite the solution to

$$\begin{aligned}
\boldsymbol{a}_0 &= \left(\mathbf{V} \boldsymbol{\Sigma}^2 \boldsymbol{V}^T + \omega^2 \boldsymbol{I} \right)^{-1} \boldsymbol{X}^T y && (4.19)\\
&= \mathbf{V} \left(\boldsymbol{\Sigma}^2 + \omega^2 \boldsymbol{I} \right)^{-1} \boldsymbol{V}^T \boldsymbol{X}^T y, && (4.20)
\end{aligned}$$

where the inverse can now easily be found by the inverse to the diagonal entries of $\boldsymbol{\Sigma}^2 + \omega^2 \boldsymbol{I}$. As it can be seen from equation 4.20 above, we only need the $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$, so if the number of data points far exceed the number of parameters ($N_d > N_p$) it will be more efficient to make the SVD of $\boldsymbol{X}^T \boldsymbol{X}$, where $\boldsymbol{V\Sigma}^2 \boldsymbol{V}^T = \boldsymbol{X}^T \boldsymbol{X}$.

Minimizing RR-LS cost function in equation 4.15 with too many parameters for $\omega = 0$ is an ill-posed problem, in which there will be zero or near zero singular values. The corresponding modes are called sloppy modes.[4] For sloppy modes an insignificant change of the data will cause a large change in the solution, and the parameters are therefore poorly determined. A larger ω will however make the problem well-defined.

The Hessian to the cost is given as

$$\boldsymbol{H} = \left. \frac{\partial^2 C(\boldsymbol{a}, \omega)}{\partial \boldsymbol{a}^2} \right|_{\boldsymbol{a}=\boldsymbol{a}_0} = \boldsymbol{X}^T \boldsymbol{X} + \omega^2 \boldsymbol{I}, \qquad (4.21)$$

and with the definition $C_0(\omega) = C(\boldsymbol{a}_0(\omega))$, we can write the cost function as

$$C(\boldsymbol{a}, \omega) = C_0(\omega) + \frac{1}{2} (\boldsymbol{a} - \boldsymbol{a}_0)^T \boldsymbol{H} (\boldsymbol{a} - \boldsymbol{a}_0), \qquad (4.22)$$

which will become useful later.

Lastly we define the smoother matrix as[5]

[4] See [Brown and Sethna, 2003]

[5] See [Hastie et al., 2009]

$$\begin{aligned} \boldsymbol{S} &= \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}+\omega^2\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T && (4.23)\\ &= \boldsymbol{X}\mathbf{V}\left(\boldsymbol{\Sigma}^2+\omega^2\boldsymbol{I}\right)^{-1}\boldsymbol{V}^T\boldsymbol{X}^T && (4.24)\\ &= \boldsymbol{U}\left(\boldsymbol{\Sigma}^2+\omega^2\boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}^2\boldsymbol{U}^T, && (4.25)\end{aligned}$$

and the model predictions for the target values can thus be given as

$$\boldsymbol{y}^{\text{model}} = \boldsymbol{X}\boldsymbol{a} = \boldsymbol{S}\boldsymbol{y},$$

which we will use in the following.

4.3.1 The effective number of parameters[6]

The regularization is smoothing out how the parameters should depend on the data, and one can therefore talk about the effective number of parameters (N_{eff}), or the effective degrees of freedom, in the model.

It is therefore very convenient to define N_{eff} by the sum of the diagonal elements in the smoothing matrix of equation 4.23, hence

$$\begin{aligned} N_{eff}(\omega^2) &= \text{tr}(\boldsymbol{S}), && (4.26)\\ &= \text{tr}(\left(\boldsymbol{\Sigma}^2+\omega^2\boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}^2), && (4.27)\\ &= \sum_p \frac{\Sigma_p^2}{\Sigma_p^2+\omega^2}, && (4.28)\end{aligned}$$

where $\text{tr}(\cdot)$ denote the trace. We can use the SVD for $\boldsymbol{X}^T\boldsymbol{X}$ also to calculate N_{eff}.

We note that the following limits are given for the effective number of parameters for regularization strength:

$$\begin{aligned} \lim_{\omega^2\to 0} N_{eff} &= \sum_p 1 = N_p, && (4.29)\\ \lim_{\omega^2\to\infty} N_{eff} &= \sum_p 0 = 0. && (4.30)\end{aligned}$$

We will use N_{eff} to describe how complex the model is when providing solutions to the cost function of different fitting problems.

4.3.2 The Bayesian error estimation ensemble[7]

Using Bayes's theorem it is possible to ask what is the model uncertainty from knowing some data $\mathcal{D}$, which is the basis for creating an error estimation ensemble for the fitted functional.

[6] See [Hastie et al., 2009]

[7] See [Brown and Sethna, 2003].

A probability distribution P can be written fluctuation $\delta\mathbf{a}$ around $\mathbf{a}_0$. For this probability distribution we will require that the mean expectation value of $\mathbf{a}' = \mathbf{a}_0 + \delta\mathbf{a}$ reproduce the mean error of $\mathbf{a}_0$, hence:

$$\sum_i^{N_d} \left\langle (\delta q_i)^2 \right\rangle = \sum_i^{N_d} (\Delta q_i)^2 \,, \tag{4.31}$$

where $\langle \ldots \rangle$ indicates the average over the ensemble, and, δq_j , and Δq_j are the prediction errors given respectively by $\mathbf{a}'$ and $\mathbf{a}_0$. In other words we demand that the error estimation prediction ensemble on average reproduce the observed error of $\boldsymbol{a}_0$. A probability distribution for the fluctuations that fulfills the above requirement is

$$\begin{aligned} P &\propto \exp\left(-C(\mathbf{a})/\tau\right), & (4.32) \\ \tau &= 2C_0/N_{eff}, & (4.33) \end{aligned}$$

where the ensemble temperature τ scales the model fluctuations in accordance to equation 4.31. The unbiased temperature τ is given by scaling the temperature with $\frac{N_d}{N_{eff}+N_d}$. Each of the parameters are assumed to contribute equally to the cost function in a harmonic fashion.

In effect of the temperature τ will scale the Hessian and we define

$$\mathbf{\Omega} = \tau \mathbf{H}^{-1}. \tag{4.34}$$

The ensemble of fluctuations can be now created using a random number generator and the eigenvalue decomposition of the $\mathbf{\Omega}$, hence

$$\delta\mathbf{a}_k{=}\mathbf{V}{\cdot}\mathrm{diag}(\sqrt{\mathbf{u}}){\cdot}\mathbf{v}_\mathbf{k}, \tag{4.35}$$

where, $\mathbf{v}_\mathbf{k}$ is a N_p long random vector of normal distributed numbers (variance 1 and mean 0); and $\mathbf{u}$, and $\mathbf{V}$ are the eigenvalues and eigenvectors of $\mathbf{\Omega}$ respectively.

The Bayesian error estimation (BEE) ensemble prediction error on data points i can therefore be calculated directly as

$$\sigma_{\text{BEE}} = \sqrt{\mathbf{x}_i \mathbf{\Omega}^{-1} \mathbf{x}_i^T} \,, \tag{4.36}$$

and the BEE ensemble prediction of the entire dataset is given by

$$\sigma_{\text{BEE}} = \sqrt{\mathbf{X}\mathbf{\Omega}^{-1}\mathbf{X}^T}.$$

If we disregarding dependency between the data points, we can define a covariance matrix from the BEE ensemble prediction as

$$\mathrm{Cov}_{\mathrm{BEE}} = \boldsymbol{I}\sigma^2_{BEE}. \tag{4.37}$$

4.3.3 *Origo of the solution*

In the ridge regression formalism given, the penalization of the solution has been done from a zero vector origo. The origo of the prior solution can however be shifted to accommodate for a a prior belief in a solution that we would use if we did not have any data at hand. With this the regularization term is defined as

$$R(\boldsymbol{a}, \boldsymbol{a_p}, \omega) = \frac{1}{2}\omega^2 \left(\boldsymbol{a} - \boldsymbol{a}_p\right)^T \left(\boldsymbol{a} - \boldsymbol{a}_p\right), \tag{4.38}$$

where $\boldsymbol{a}_p$ is the prior origo of the solution. To bring regularization back to the ridge regression form, this prior solution shift is transformed into the target vector $\boldsymbol{y}$ following

$$\mathring{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}_p, \tag{4.39}$$

where $\mathring{\boldsymbol{y}}$ is then used in the cost function instead of $\boldsymbol{y}$, and the solution coefficient vector is afterwards adjusted back to the zero origo reference model space by

$$\boldsymbol{a} = \mathring{\boldsymbol{a}} + \boldsymbol{a}_p. \tag{4.40}$$

4.3.4 *Tikhonov regularization*

The ridge regression treats all the different parameters equally, however at times the different parameters can have quite a different effect - e.g. for a polynomial base the 1st order polynomial and the 10th order polynomial will affect model prediction quite differently. We therefore generalize the ridge regression to the Tikhonov regularization form, where the Tikhonov matrix $\boldsymbol{\Gamma}$ is introduced to govern the regularizing scaling between the parameters, hence

$$R(\boldsymbol{a}, \omega) = \frac{1}{2}\omega^2 \boldsymbol{a}^T \boldsymbol{\Gamma}^2 \boldsymbol{a}, \tag{4.41}$$

where ridge regression corresponds to the case $\boldsymbol{\Gamma} = \boldsymbol{I}$.

A commonly used choice of the Tikhonov matrix is to penalize non-smoothness of the underlying function, hence

$$\boldsymbol{\Gamma}^2_{ij} = \int \frac{\partial^2 f_i(x)}{\partial x^2} \frac{\partial^2 f_j(x)}{\partial x^2} dx. \tag{4.42}$$

To make $\boldsymbol{\Gamma}$ invertible however, we need to add a diagonal elements to overlap entries that are zero. This is usually done by adding a diagonal constant matrix to the Tikhonov matrix, hence

$$\boldsymbol{\Gamma}_{ij}^2 = \int \frac{\partial^2 f_i(x)}{\partial x^2} \frac{\partial^2 f_j(x)}{\partial x^2} dx + \alpha \boldsymbol{I}, \tag{4.43}$$

where α is some small constant that makes the inversion well-defined. Using that the Tikhonov matrix is invertible, a transformation of the problem can be made such that the ridge regression form reappear. By the transformations

$$\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{\Gamma}^{-1} \quad \boldsymbol{y}' = \boldsymbol{y} \quad \boldsymbol{a}' = \boldsymbol{\Gamma}\boldsymbol{a}, \tag{4.44}$$

one can now solve the ridge regression problem as before but in the prime space. The solution is then given in the original "non-smooth" parameter space through the transformation $\boldsymbol{a} = \boldsymbol{\Gamma}^{-1}\boldsymbol{a}'$.

It should be noted that the smoothness can be multidimensional, and it is then given by the laplacian operator.

Transforming the input space with a Padé approximant will result in a different definition of smoothness for the problem, when the smoothness is defined in the transformed space. The transformation therefore serves a double role by making the it possible to describe the underlying model better with fewer parameters, and by allowing a the flexibility of the model where it is judged to be important with the smoothness regularization.

In the following we will investigate how to decide on the optimal regularization strength, i.e. choosing the optimal model complexity.

4.4 *Optimal model complexity.*

For deciding the optimal model complexity, we will use cross validation, where we optimize for the model complexity that yields the most transferable model.

4.4.1 *Cross-validation*

The optimal model complexity has to rely on the available data and our prior knowledge, but we need a way to figure out if the model is an overfit for selecting the regularization strength. For every model optimization we can divide the data into 4 groups:

Known known the training data

Known unknown data we know about but don't use

Unknown known the selection of the data or handed over implicit knowledge about how the data should behave

Unknown unknown the data that we would like the model to work on in the end

The optimal model is one that is able to describe the unknown unknown data, and we are going to use cross-validation to give an estimate that will lead us to choose the model that is hopefully best suitable for this.

We could divide the data in two groups: a training set (known known) and a holdout set (known unknown) for validating the model. We can use the validation data to test for overfitting, hence transferability of the fitted model. We would however like to use all the data to train our model.

In cross-validation one rotate the data between the holdout and training group, and use evaluation on the validation data to select the optimal regularization strength for the full dataset. Commonly used cross-validation are of the group K-fold cross validation, where $1/K$ of the data is used as holdout data, and the rest training data. The data can be taken out either randomly or in a rotation. Ideally one would exhaust the number of ways to draw the holdout data but that is computationally unfeasible.

The Leave-one-out technique, which name describes the method, is of the rotation type, where one can do an exhausted sampling, and Bootstrapping one will on the contrary randomly pick the validation data. Of popular K-fold method that will not be presented in more depth are 10 fold, were one randomly pick $1/10$ of the validation data in each round and fit to the rest, and the 2 fold method where one takes half the data out and the make a cross examination by training on one group and validating on the other and vice versa.[8]

4.4.2 *Leave-One-Out Cross-Validation*

The leave one out cross-validation (LOOCV) is a very simple cross-validation technique for which a very fast solution exists. The estimated prediction error of the method is the average of the validation error when taking one data point out of the training dataset as validation data. In practice a closed form solution exists such that one does not need to do N_d minimizations of the cost function. For a dataset with a set data points given by the set $\mathcal{Z}$.

Let the leave-one-out cross validation loss estimation be given as

$$\hat{\sigma}^2_{LOO} = \frac{1}{N_d} \sum_{i=1}^{N_d} L_i(\boldsymbol{a}_{i \notin \mathcal{Z}}), \tag{4.45}$$

where $\boldsymbol{a}_{i \notin \mathcal{Z}}$ is the solution to the full cost function for which data point i has been omitted from the dataset $\mathcal{Z}$. The fast LOOCV esti-

[8] See [Hastie et al., 2009]

mate will be given for the LS-RR cost function, where the models are found by LS-RR and the loss function for the estimator is that of LS.

Define

$$\boldsymbol{P} = I_{N_d \times N_d} - \boldsymbol{S}, \tag{4.46}$$

where $\boldsymbol{S}$ is the smoothing matrix from equation 4.23. The leave-one-out least squares prediction error can then be found by the following

$$\hat{\sigma}^2_{LOO-LS} = \frac{1}{N_d}\boldsymbol{y}^T\boldsymbol{P}\left(diag(\boldsymbol{P})\right)^{-2}\boldsymbol{P}\boldsymbol{y}. \tag{4.47}$$

We will in the following refer to the least square LOOCV as LOOCV or LOO.

4.4.3 *Bootstrap resampling and the Bootstrap .632 estimate*

With bootstrap resampling a number of new datasets are created from the original with random picking from the original dataset, where a data point is allowed to be used several times. The statistical measures are then calculated from looking at the distribution of all the samples. The bootstrap .632 estimator[9] is based on the observation that using some of the data as validation data and not omitting it as training dataset a less than optimal model is created. The true error of the trained model should be expected to be better as more data is used in training the final model than in the cross-validation estimations. The bootstrap 0.632 estimator corrects for this by mixing the model variance to the full dataset with the estimated error variance from the bootstrap cross-validation $\widehat{ERR}$ in the following manner:

$$\text{EPE}_{.632} = \sqrt{0.632 \cdot \widehat{\text{ERR}} + 0.368 \cdot \widehat{\text{err}}}. \tag{4.48}$$

The estimate of $\widehat{\text{err}}$ is the error estimation by the best fit to the entire dataset, given as

$$\widehat{\text{err}} = L(a_0) \tag{4.49}$$

and $\widehat{\text{ERR}}$ is the bootstrap estimation of the error defined by

$$\widehat{\text{ERR}} = \frac{1}{N_d}\sum_i \frac{1}{N_{s|j\notin s}} \sum_{j=s|j\notin s} L_i(\mathbf{b}_j), \tag{4.50}$$

where with N_s bootstrap samples for each data point i in the dataset there are $N_{s|j\notin s}$ samples where the data point is not a part of the sample training set s. For each datapoint the loss function for that specific datapoint is used, and $\boldsymbol{b}_i$ is the parameters that minimizes the model cost function C_s given for the training set. Following the discussion before, each sample s is created by resampling of the

[9] [Efron, 1983]

dataset, where each data point is allowed to be used multiple times in the sample dataset.

Using the same samples for each regularization strength removes sample variation noise in the EPE regularization curve. It is important that enough samples are being used to overcome the noise of the method.

For the RR-LS cost function with the use of the same samples for all regularization strengths, we can made the SVD for all samples squared design matrices, $\boldsymbol{X}^T\boldsymbol{X}$ once and thus calculating the bootstrap error for different regularization strengths at almost no cost.

4.5 A geometric mean cost and loss function

When the model needs to be optimized for several properties that are not directly comparable, the arithmetic mean (μ_A), $\mu_A = \frac{1}{n}\sum_i^n x_i$, might not be ideal. For example for comparing two models on two clusters, where the target data values of one is magnitudes larger than the other; here the cluster with smaller values be overshadowed by the one with bigger values. The scaling of the two clusters might be arbitrary, and the best model is therefore not properly defined.

An alternative to the arithmetic mean is the geometric mean (μ_G or GM), $\mu_G = \left(\prod_i^n x_i\right)^{1/n}$; here the scale of the different properties do not matter. When comparing two models by the geometric mean, if one model 1 can lower x_1 by 10 percent while only making x_2 higher by less than 10 percent compared to model 2, then model 1 will have a smaller geometric mean assuming all other things being equal.

To find model compromises between several clusters, the geometric mean will therefore be used in different forms. The estimate can however be used in different ways, and this compromise estimate furthermore has consequences for also the model selection.

The training data $\mathcal{D}$ is divided into N_k clusters, and $\mathcal{D}_k$ is now the data points associated with cluster k. Note that the clusters can have overlap between them. In the following the cost function C_k and loss function L_k are the cost and loss function for the data points of cluster k, i.e. $\mathcal{D}_k$.

4.5.1 Geometric mean of cluster cost functions

A simple model compromise when fitting several clusters is that of the geometric mean of each clusters' cost functions, where each clusters' cost function includes the optimal regularization strength found for that cluster. The global cost for the N_k clusters is then given by

$$\breve{C}_{GM} = \left(\prod_{k}^{N_k} C_k^{\breve{w}_k} \right)^{1/\sum \breve{w}_k}, \tag{4.51}$$

where $\breve{w}_k$ is the weight of cluster k, and C_k is the cost function to the individual cluster with the minimum solutions $\boldsymbol{a}_k^*$. By adding a weight to the individual clusters one has the ability to control for for instance if two clusters describe the same property, or if one property is more important than the other.

This cost function of equation 4.51 is non-linear when $N_k > 1$, and a new approach to solving the global cost function is therefore needed. For the following it is assumed that the cost functions of the individual clusters are harmonic, e.g. RR-LS cost functions.

One can start by defining the cost that we want to solve in terms of the logarithm to the cost, hence

$$\breve{K}_{GM} = \ln\ \breve{C}_{GM} = \sum_{k}^{N_k} \frac{\breve{w}_k}{\sum \breve{w}_k} \ln\ C_k, \tag{4.52}$$

and using the zero gradient condition for the minimum we have

$$\sum_{k}^{N_k} \frac{\breve{w}_k}{\sum \breve{w}_k} \frac{1}{C_k} \frac{dC_k}{d\boldsymbol{a}} = 0. \tag{4.53}$$

Using the Hessians and individual solutions of the cost functions we can write

$$\sum_{k}^{N_k} \frac{\breve{w}_k}{\sum \breve{w}_k} \frac{1}{C_k} \frac{d}{d\boldsymbol{a}} \left(C_k(\boldsymbol{a}_k^*) + \frac{1}{2} \left(\boldsymbol{a} - \boldsymbol{a}_k^*\right)^T \boldsymbol{H}_k \left(\boldsymbol{a} - \boldsymbol{a}_k^*\right) \right) = 0, \tag{4.54}$$

and with $\mathcal{W}_k = \frac{\breve{w}_k}{\sum \breve{w}_k} \frac{1}{C_k}$, the solution is given by

$$\breve{\boldsymbol{a}} = \left(\sum_{k}^{N_k} \mathcal{W}_k \boldsymbol{H}_k \right) \left(\sum_{k}^{N_k} \mathcal{W}_k \boldsymbol{H}_k \boldsymbol{a}_k^* \right). \tag{4.55}$$

$\mathcal{W}_k$ is however not known initially, but can be approximated iteratively. Starting from an initial guess to the solution, e.g. $\breve{\boldsymbol{a}}_{init} = \frac{1}{N_k} \sum \boldsymbol{a}_k^*$, one calculated $\mathcal{W}_k$ which lead to a new guess of $\breve{\boldsymbol{a}}$, and the process repeats.

The geometric mean of the cost functions provides a well-defined direct compromise. The added knowledge from accumulating data of several clusters can however not be used to create a more complex model.

4.5.2 Geometric mean of cluster loss functions

When more consistent data is at hand, we should be confidence about finding a more complex solution, without having to fear that we are overfitting. This suggest that the loss function should be the geometric mean of cluster loss functions, and we the full cost function can be given in the form

$$\check{C}(\boldsymbol{a},\omega) = \left(\prod_{k}^{N_k} L_k(\mathbf{a})^{\check{\mathrm{w}}_\mathrm{k}}\right)^{1/\sum \check{w}_k} e^{R(\mathbf{a};\omega)}, \tag{4.56}$$

where as previous L_k is the loss function for cluster k, and $\check{w}_k$ is an adjustable weighting of each cluster. The loss function L_k is taken to be the average loss of the dataset, such that the geometric loss function is consistent for different sizes of datasets.

The logarithm to the cost function of equation yields

$$K(\mathbf{a};\omega) = \ln\{\check{C}(\boldsymbol{a},\omega)\} = \sum_k \frac{\check{w}_k}{\sum \check{w}_k} \ln\{L_k(\mathbf{a})\} + R(\mathbf{a};\omega). \tag{4.57}$$

Using the zero gradient condition to find the minimum of the cost function yields

$$\frac{\partial K}{\partial \mathbf{a}} = \mathbf{0} = \sum_k \frac{w_k}{N_k} \frac{\partial \ln L_k}{\partial \mathbf{a}} + \frac{\partial R}{\partial \mathbf{a}} = \sum_k \frac{w_k}{L_k \sum \check{w}_k} \frac{\partial L_k}{\partial \mathbf{a}} + \frac{\partial R}{\partial \mathbf{a}}, \tag{4.58}$$

which is a again a non-linear problem.

Equation 4.58 can either be solved through an iterative procedure, were it is assumed that the loss function is quadratic. A new cost function with the same minimum as equation 4.56 is defined by

$$\begin{aligned} \tilde{K}(\boldsymbol{a};\omega,\boldsymbol{a}_*) &= \tilde{L}_k(\boldsymbol{a};\boldsymbol{a}_*) + R(\boldsymbol{a};\omega) & (4.59)\\ &= \sum_k \frac{w_k}{\sum w_k} \frac{1}{L_k(\boldsymbol{a}_*)} L_k(\boldsymbol{a}) + R(\boldsymbol{a};\omega) & (4.60) \end{aligned}$$

where $\boldsymbol{a}_*$ is the best guess of the solution. $\tilde{K}$ can now solved, starting with an initial solution guess $\tilde{\boldsymbol{a}}_{init}$: $\boldsymbol{a}_*$ is iteratively replaced with $\tilde{\boldsymbol{a}}$ and convergence is reached when minimizing $\tilde{K}$ yields the self-consistent solution $\boldsymbol{a} = \boldsymbol{a}_*$. The converged solution of $\tilde{K}$ is denoted $\tilde{\boldsymbol{a}}$. $\tilde{K}(\boldsymbol{a};\omega,\boldsymbol{a}_0)$ is quadratic, and the Hessian can therefore be found in the usual manner. The same goes for the number of effective parameters (using equation 4.26) and the Bayesian error estimation ensemble; all given for the cost function $\tilde{K}(\boldsymbol{a};\omega,\tilde{\boldsymbol{a}})$.

It is also possible to minimize the non-linear cost function K directly with BFGS and $\tilde{K}$ is then given from the solution found by BFGS.

4.5.3 Cluster-LOOCV for model compromise

We define the cluster-leave-out-out cross validation as by rotation of each cluster as holdout set and then evaluate the fit from the other clusters on it.

Two alternative prediction errors are given for cluster-LOOCV, that of the arithmetic mean of the validation loss and the geometric mean of the validation loss. Let Z be the ensemble of clusters, i.e. $Z = \{k_1, k_2, k_3, \dots, k_{N_k}\}$. The arithmetic mean estimation is now given as:

$$\Delta^2_{AM} = \frac{1}{N_k} \sum_k L_k(\tilde{\boldsymbol{a}}_{k \cap Z}), \tag{4.61}$$

and the geometric mean of validation loss

$$\Delta^2_{GM} = \left(\prod_k L_k(\tilde{\boldsymbol{a}}_{k \cap Z})^{w_k} \right)^{1/\sum w_k}, \tag{4.62}$$

where the loss function is similar to the one optimized for the individual clusters, hence for one dataset. $k \cap Z$ is the clusters excluding the k'th cluster.

4.5.4 Hierarchical cluster-Bootstrap.632

The geometric mean loss function is now introduced in the bootstrapping formalism. As previously, the prediction error in the Bootstrap$_{.632}$ formalism is given as $EPE_{.632} = \sqrt{0.368\widehat{err} + 0.632\widehat{ERR}}$, see equation 4.48-4.50. In a geometric mean cluster version we define

$$\widehat{\mathrm{err}} = \left(\prod_k^{N_k} L_k(\tilde{\boldsymbol{a}})^{w_k} \right)^{1/\sum \mathrm{w}_k}, \tag{4.63}$$

where L_k is again the loss function corresponding to the individual clusters, hence the average squared loss. Further we define

$$\widehat{\mathrm{ERR}} = \left(\prod_k^{N} \left(\frac{1}{N_{\mu_k}} \sum_{\mu_k} L_k(\boldsymbol{b}_{\mu_k}) \right)^{\mathrm{w}_k} \right)^{\frac{1}{\sum \mathrm{w}_k}}, \tag{4.64}$$

where μ_k iterates over bootstrap samples s, where cluster k has been omitted from the training set. N_{μ_k} is the number of bootstrap samples where cluster k is not in the sample. And $\boldsymbol{b}_{\mu_k}$ is the solution to the μ_k'th global cost function, hence

$$\mathbf{b}_{\mu_k} = \underset{\mathbf{b}}{argmin} \left(\breve{C}_{\mu_k}(\boldsymbol{b}, \omega) \right).$$

$\breve{C}_{\mu_k}$ is the global cost function of equation 4.56 for the μ_k'th sample of the bootstrap resampled training data, i.e. s_{μ_k}.

The resampling procedure now takes a hierarchical form: For each sample s each cluster is bootstrap resampled internally. Secondly, the collection of resampled clusters in the sample is bootstrap resampled.

The rationale for this two level resampling scheme is to have a method that will capture correlations between the data points in the clusters and between the clusters. In other words we want the procedure to capture if two clusters are highly correlated, but also if the clusters are internally correlated.

With the normal Bootstrapping method it is unlikely that a large cluster will be taken out in one of the sample, for a limited number of bootstrap samples, and so the method will not be able to test transferability between the different clusters fully. The internal resampling makes the method resilient to the case where the clusters are a copies of one another.

4.6 Robust fitting

The downside of using the least squares as loss function, which we have discussed up until now, is that if one data point is very off, i.e. an outlier, then the whole model will break down and not be transferable. This is the price that is paid from from having highest possible efficiency estimator for is normally distributed data.

The median loss function estimator, on the other hand, has a breakdown proportion of 0.5, as half of the data points can be outliers without it affects the estimate of the median significantly. The median however has a low efficiency on normally distributed data compared to the least squares. Ideally we would like to use an estimator that simultaneously has a high efficiency and a high breakdown point. The loss functions of robust statistics seeks exactly that.

An outlier can have a big influence on the predicted model when using a non-robust estimator, and thus creating a masking effect. For this reason one cannot rely on inspection to identify outliers, and thus removing them manually.

4.6.1 M-estimators

The M-estimator is a generalization of the maximum likelihood estimator.[10] Given a linear fitting problem, the residual vector for a model $\mathcal{M}$ is given by $\boldsymbol{r} = \boldsymbol{Xa} - \boldsymbol{y} = (r_1, ..., r_n)^T$, with the notation of previous sections. The M-estimator is now defined the solution of $\hat{\sigma}$ in

$$\frac{1}{N_d} \sum_{i=1}^{n} \rho \left(\frac{r_i}{\hat{\sigma}} \right) = \delta, \tag{4.65}$$

where $\delta \in [0, 1]$ and ρ is a ρ-function. The solution $\hat{\sigma}$ is called the M-estimator of scale. Take the example of the least squares ρ-function $\rho(t) = t^2$, which we saw earlier was associated with a gaussian error probability distribution. The solution to the M-estimator of scale for $\rho(t) = t^2$ and $\delta = 1$, is using equation 4.65(above) $\hat{\sigma} = \sqrt{ave(\boldsymbol{x}^2)}$, thus yielding the root mean square (RMS). In other words the RMS is the M-scale of the least squares loss function. Now by choosing a different ρ-function and tuning the δ value, one can find a more robust estimation of the scale.

The M-estimator can be used for regression by using the ρ-function as the loss function, with every data point scaled by the M-scale of equation 4.65.[11]

[10] See [Huber, 1964]

[11] See [Maronna et al., 2006]

The δ has the role of determining the breakdown point of the estimator. The breakdown point (BDP) is mathematically defined as the maximum proportion of observations that can be arbitrary altered with the estimator remaining bounded away from the border or the parameter set.

A ρ-function, $\rho(t)$, is a nondecreasing function of $|t|$, where $\rho(0) = 0$, $\rho(t)$ is increasing for $t > 0$ such that $\rho(t) < \rho(\infty)$, and, if ρ is bounded, $\rho(\infty) = 1$. A number of robust ρ-functions have been suggested. For these the general property is ρ asymptotically becomes flat such that big outliers will not influence the estimate more than small outliers. It ρ is differentially for all t, it is called a ψ-function.

A frequently used ρ-function for scale estimations is the Tukey-Bisquare defined as

$$\rho_{bis}(t) = \min\{1 - (1 - t^2)^3, 1\}, \tag{4.66}$$

which is compared to least squares in the figure to the right.
The TukeyBisquare is a ψ-function.

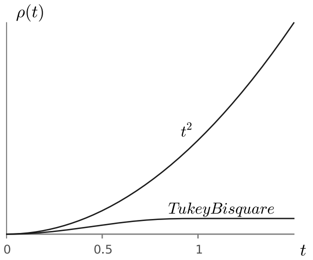


Figure 4.2: Comparing the Tukey-Bisquare and least squares ρ-functions.

To find the scale for a set of residuals, numerically procedures can be used. Using those, the next section will let us solve the linear regression problem for a the ridge regression problem with a ψ-function loss function.

4.6.2 The iterative reweighting least squares (IRWLS)

To solve the regression problem with the M estimator, one can use the Iterative ReWeighting Least Squares (IRWLS), starting from good guesses to the final solution. The approach uses that for the RR-LS cost function, the solution could be found in a simple close form: $\boldsymbol{a} = \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{I}_M\omega^2\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$, see section 4.3. For the M estimator with ridge regression, i.e. $L = \rho(t)$, a similar system of equations can be created.[12] We define

$$\psi(t) = \rho'(t), \qquad W(t) = \frac{\psi(t)}{t}, \tag{4.67}$$

and

$$t_i = \frac{r_i}{\hat{\sigma}}, \qquad w_i = \frac{W(t_i)}{2} \tag{4.68}$$

$$\boldsymbol{w} = (w_1, \ldots, w_n)', \qquad \boldsymbol{W} = diag(\boldsymbol{w}). \tag{4.69}$$

Now differentiating the M-estimator cost function with the solution vector, $\boldsymbol{a}$, and setting it equal to zero, and a similar solution to RR-LS appears, hence

[12] See [Maronna, 2011] for ridge regression solution and [Maronna et al., 2006] for the M-estimator solution.

$$\boldsymbol{a} = \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} + \boldsymbol{I}_M \omega^2\right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y}, \tag{4.70}$$

which is a weighted version of the solution to the RR-LS solution. Since ρ and thus $W(t)$ are decreasing functions of $|t|$, observations with large residuals will receive lower weights w_i. The weighted normal equations suggest an iterative procedure:

1. From an initial solution $\boldsymbol{a}_0$ calculate the M-estimator: , $\hat{\sigma}_0$.
2. For $k = 0, 1, 2, ...$:
   (a) Given the solution $\boldsymbol{a}_k$: calculate the residual vector $\boldsymbol{r}$, the M-estimator $\hat{\sigma}_k$ and then the weights $\boldsymbol{W}$
   (b) Compute $\boldsymbol{a}_{k+1}$ by solving 4.70.
3. Stop if the residuals changes are below a predefined threshold.

The iterative reweighting procedure to solve the ridge regression M-estimator problem has in [Maronna, 2011] been found to decrease the object function of the cost function at every step, whereas the solution to the normal M-estimator problem, i.e. without regularization, with iterative reweighting least squares has been proven to decrease at every iteration. The initial solution should have a high breakdown point and will as a consequence possible have a low efficiency.

As the loss function is no longer quadratic, many local minima are possible. Therefore one have to rely a good initial solutions to start the IRWLS, or an ensemble of initial solutions. The initial solutions can be found using various scale independent estimator. e.g. the $L = |\boldsymbol{X}\boldsymbol{a} - \boldsymbol{y}|$, or by trimming the dataset in different ways and solve for the RR-LS.

The weighting of the data points means that we have to reestablish the effective number of parameters for a given regularization strength. Using the definition of $\boldsymbol{w}$ from equations 4.68 and 4.69 and the following transformations:

$$\tilde{x} = \frac{\boldsymbol{w}'\boldsymbol{X}}{\boldsymbol{w}'\boldsymbol{1}}, \qquad \underline{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{1}\tilde{x} \tag{4.71}$$

The Hessian for the reweighed calculations is given as

$$\hat{\boldsymbol{H}} = \underline{\boldsymbol{X}}(\underline{\boldsymbol{X}}'\boldsymbol{W}\underline{\boldsymbol{X}} + \omega^2 \boldsymbol{I}_M)^{-1}\underline{\boldsymbol{X}}'\boldsymbol{W} \tag{4.72}$$

Again following the procedure in section4.3.1 and the definition of [Hastie et al., 2009], effective degrees of freedom and number of effective parameters are given as

$$\hat{N}_{eff} = \operatorname{tr}\left(\hat{\boldsymbol{H}}\right). \tag{4.73}$$

Following, N_{eff} will refer to $\hat{N}_{eff}$ for MM calculations.

4.6.3 The MM-estimator

The MM-estimator is an extension to the M-estimator, where a two step process is used of to find the optimal solution and achieve a higher breakdown point and efficiency simultaneously. Two different ρ-functions are used: and initial function ρ_0 for high breakdown point(BDP) and a function ρ for the efficient of the estimate. Both are bounded and $\rho_0 \geq \rho$.[13]

It can be shown that if $L(\boldsymbol{a}_{\rho_0}) \geq L(\boldsymbol{a}_{\rho})$ then the solution is consistent, and the BDP of ρ is no less than of ρ_0. Furthermore if ρ is differentiable, then a solution to the cost function will have the same efficiency as the global solution. Using these properties one can find a solution with high efficiency and sufficient breakdown by using a ρ_0 for high BDP and ρ for a sufficient efficiency.

We start with an initial estimator $\boldsymbol{a}_{init}$, from which we calculate the residuals $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{a}_{init})$, and an M-scale $\hat{\sigma}_{init}$ given by

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{r_i}{\hat{\sigma}_{init}}\right) = \delta. \tag{4.74}$$

The cost function for the MM estimator for ridge regression (RR-MM) is now given in the form

$$C(\hat{\boldsymbol{a}},\omega) = \hat{\sigma}_{init}^2 \sum_{i}^{N_d}\rho\left(\frac{r_j(\hat{\boldsymbol{a}})}{\hat{\sigma}_{init}}\right) + R(\boldsymbol{a},\omega), \tag{4.75}$$

where the factor $\hat{\sigma}_{init}^2$ in front the estimator is added to make the loss function coincide with the least squares cost function one for $\rho(t) = t^2$. The minimization problem is solved with IRWLS procedure as shown above.

[13] See [Yohai, 1987]

5 Bayesian Error Estimation functionals

The studies BEE functional previous my involvement will will here be presented.

5.1 Inspiration for density functionals with error estimation

The development of Kohn-Sham density functionals with error estimation was inspired by the development of interatomic potentials with error estimation.[1] Interatomic potentials are used extensively to study the structural and dynamical properties of a wide range of materials from *biomolecules* to polymers and semiconductors. Typically interatomic potentials are computational very fast, because the interatomic energies and forces are carried out by explicit evaluations of pair-like or angular terms, that depends on only the coordination of a few atoms at a time. It is therefore possible to look at much larger systems than DFT. The speed and simplicity of the interatomic potentials are, however, obtained at a cost of lower prediction power, and the accuracy of the potentials is therefore a concern.

The study of Frederiksen et al. [2004] therefore presented a method for how to create a Bayesian error estimation ensemble to assess the uncertainty of the degree of uncertainty that should be expected using the potentials. The method employs a harmonic approximation of a cost function to construct the ensemble in the fashion outlined in the machine learning introduction.[2] The cost function for the parametrization was the squared least between between the model and the DFT references. To ensure transferability, and avoid overfitting, a holdout set is used to check how many parameters should be used in the potential.

This study thus lead the way to how the Bayesian error estimation could be used for fitted models within electronic structure calculations. Following will be presented the insights gained in the first studies on creating density functionals with error ensemble, leading up to the BEEF functional family.

[1] See Frederiksen et al. [2004]

[2] Note that for the ensemble, anharmonic effects was observed and the ensemble temperature was therefore lowered to $T_0/4$, which ensured a better agreement with the gaussian distributed reference.

5.2 *The first density functional with Bayesian error estimation*

For the first investigation of a density functional with Bayesian error estimation capabilities, it was chosen to make a parametrization of the GGA exchange enhancement factor. The parameters were then fitted to a small datasets of experimental molecular atomization energies, and experimental cohesive energies.[3] The model space expansion using the following series, $F_m(s) = \left(\frac{s}{s+1}\right)^m$, which had been chosen so that the PBE and RPBE functionals could be reproduced in the model space with few parameters in the relevant s-range.[4]

The parametrization took the following form

$$F_x^{\beta}(s) = \sum_{m=1}^{M_p} \beta_m \left(\frac{s}{s+1}\right)^{2(m-1)} , \qquad (5.1)$$

where m is the parameter, M_p is the total number of basis functions and β_m can be regarded as free parameters. The 10 lowest order basis functions are depicted in figure 5.1.

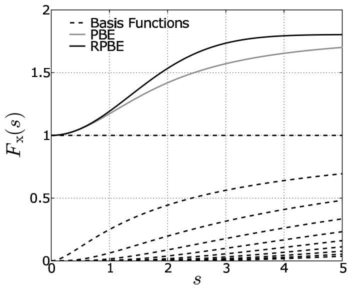


Figure 5.1: The 10 first basis functions, see Kaasbjerg [2005].

To find the optimal model the standard least square cost function was chosen: $c(\boldsymbol{\beta}) = \frac{1}{2}\sum_k \left(E_k(\boldsymbol{\beta}) - E_k^{exp}\right)^2$, where $E_k(\boldsymbol{\beta})$ is the atomization or cohesive energy of system k in the database calculated for parameters $\boldsymbol{\beta}$. The electron density is that of self-consistent PBE calculation.[5] The leave-one-out cross validation error estimate was reported to be minimized for $M = 3$ parameters in the model, and the resulting enhancement factor can be seen along with a corresponding error ensemble in figure 5.2.

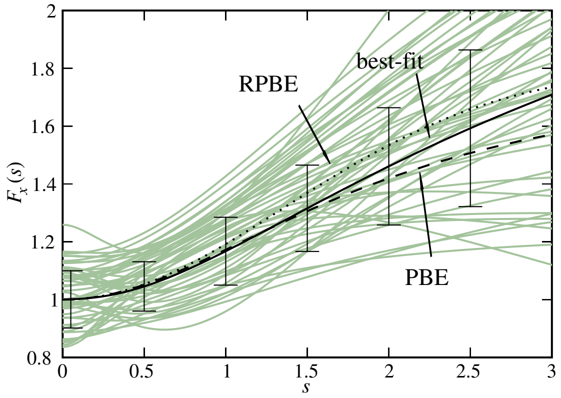


Figure 5.2: The exchange enhancement factors as a function of the dimensionless electron density gradient, adapted from Mortensen et al. [2005]. The green/grey lines show enhancement factors drawn from the error estimation ensemble of $\exp(-C(\beta)/T)$. The dashed, dotted and full lines show enhancement factors of RPBE, PBE and the best fit respectively.

The study highlighted how the ensemble could be used to judge scientific conclusions. For example what preferred binding site for a molecule is, and the uncertainty on the cohesive energy, see figure 5.3.

With this study a clear case were made for density functionals with Bayesian Error estimation; in terms of providing a simple but

[3] See Mortensen et al. [2005] and Kaasbjerg [2005].

[4] Respectively 5 and 6 parameters are need for reproducing PBE and RPBE in the interval s ∈ [0,3], see Kaasbjerg [2005]

[5] To the first order it was expected that the density would not change significant for the best fit exchange model from PBE.

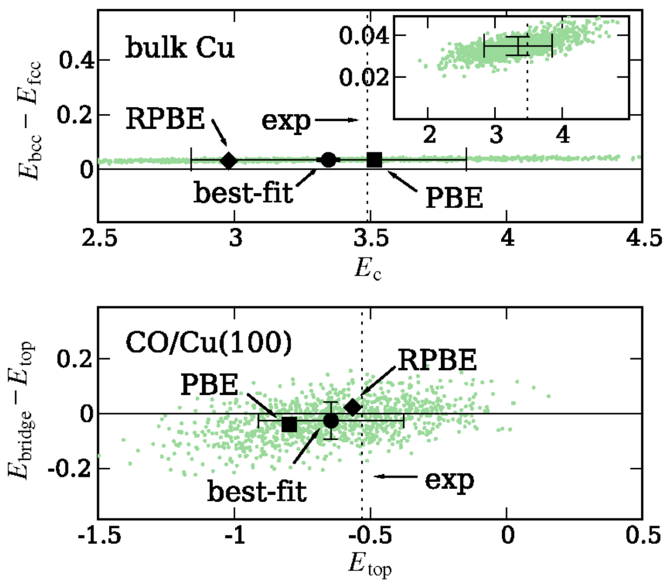


Figure 5.3: Calculated ensemble for cohesive energy (x axis) and bcc-fcc energy difference (y axis) for a copper crystal. The BEE's are indicated by error bars. The inset uses rescaled axes. Lower panel: Calculated ensemble for binding energy (x-axis) and bridge-top energy difference for CO on a CU(100) surface. Values for the experimentally preferred states (fcc and top) are indicated by vertical dotted lines. Units are in eV.

useful measure for the uncertainty to be expected when performing a DFT calculation with a GGA level exchange correlation functional.

A number of limitations in terms of the functionals model selection procedure and performance will here be highlighted. The full extend of the performance, it has later been found, cannot be assessed in non-self-consistent calculations, which will be shown later in this thesis. In the leading work of Kaasbjerg [2005] a number of issues arose that were not explicitly addressed in proceeding publication of Mortensen et al. [2005], these are highlighted in the following. In Kaasbjerg [2005] the optimal model was found by leave-one-out cross-validation to have $M = 4$ parameters, however this solution was disregarded as it was not believed to be a transferable model solution. For the cost function a suggested weighting of the systems of different materials properties were suggested, to ensure that one material property would not overrule the other in the cost function by introduce a normalization that would make the data dimensionless and normalize according to how well one regarded the data. In Mortensen et al. [2005] all data were regarded equally without taken into account of their different dimensions, and this could thus unintentionally have introduced a skewed bias towards optimizing for one of material properties fitted. In the parametrization the LDA limit were furthermore fixed and the limited flexibility of this choice towards providing a better performing functional versus having the functional be transferable outside the training data were not addressed. In following studies these potential limitations were investigated. It should also be noted that the datasets were limited in their coverage, e.g. there were no chemisorption systems in the training datasets.

5.3 *Investigation of the constraints in the GGA formalism*

In the next study a different approach were taken to the model space, which provided an easy framework for testing how different exact constraints influenced the performance of the optimal fit.[6] In the study the goal is to create a functional better suited for chemisorption, which implied that the functional would not be fitted to solid state properties. A Padé approximant was introduced as the parametrization, which took the following form:

$$F_x^{Padé}(\boldsymbol{\beta}, s) = \frac{1 + (\beta_1 + \mu)s^2 + (1+\kappa)\beta_2 s^4}{1 + \beta_1 s^2 + \beta_2 s^4}, \tag{5.2}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2)$ are the optimization parameters. In this form of the GGA exchange enhancement factor could be explored with only two parameters, and the different theoretical limits of the GGA exchange enhancement factor could be constrained or not. With this parametrization it was easy to test the implication of omitting different exact constraints on the exchange enhancement factor. These are:

(a) The local Lieb-Oxford bound: $\lim_{s\to\infty} F_x(s) = 1 + \kappa$

(b) The curvature of the exchange enhancement factor in the homogeneous electron gas limit: $\frac{\partial F_x}{\partial s^2}|_{s=0} = 2\mu$

(c) The local density limit: $F_x(s = 0) = 1$.

Using a least squared fitting procedure on a dataset of small molecule atomization energies, given in table 5.1, it was explored what omitting these constrains would impact performance. It was found that Lieb-Oxford and the gradient in the homogeneous gas limit had only minor effect on the performance (0-10% lower Mean abs. error), while the effect of omitting the the local density limit lowered the mean absolute error by 28%.

In this functional form the number of parameters are very limited, so a cross validation method is not needed. To make the ensemble estimation in the original form a harmonic cost function is required, and methods for finding a good approximation was developed as well.

A limitation of this method is that the Padé approximant model with only two parameters will expand the enhancement factor space in a certain way, and if another functional form is needed for some reason, it will not be detected in this functional space.

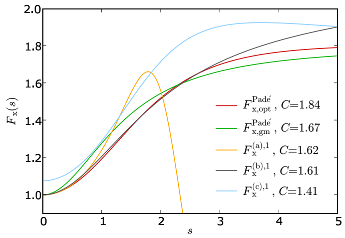


Figure 5.4: The five exchange enhancement functions used to calculate the atomization energies in table 5.1. The C values list the cost value of that functional. Figure from Toftelund [2006].

Molecule	Expt.	$F_{x,opt}^{Padé}$	$F_{x,gm}^{Padé}$	$F_x^{(a),1}$	$F_x^{(b),1}$	$F_x^{(c),1}$
H_2	4.75	4.57	4.56	4.50	4.57	4.68
LiH	2.51	2.26	2.43	2.42	2.28	2.35
CH_4	18.18	17.85	17.82	17.91	17.87	18.17
NH_3	12.9	12.74	12.82	12.81	12.62	12.83
OH	4.61	4.55	4.58	4.51	4.56	4.56
H_2O	10.07	9.71	9.64	9.59	9.72	9.77
HF	6.11	5.77	5.70	5.68	5.78	5.83
Li_2	1.06	0.87	0.90	1.14	0.85	0.90
LiF	6.02	5.18	5.50	5.65	5.24	5.36
Be_2	0.13	0.32	0.36	0.58	0.33	0.33
C_2H_2	17.58	17.49	17.53	17.44	17.53	17.64
C_2H_4	24.04	24.16	24.13	24.27	24.19	24.48
HCN	13.53	13.72	13.90	13.68	13.63	13.67
CO	11.24	11.05	11.06	10.92	11.08	10.98
N_2	9.91	10.17	10.45	10.14	9.93	9.91
NO	6.63	6.90	6.95	6.83	6.78	6.64
O_2	5.23	5.52	5.29	5.37	5.50	5.29
F_2	1.67	2.16	2.07	2.18	2.17	1.94
P_2	5.09	4.92	4.93	4.76	4.99	4.69
Cl_2	2.52	2.56	2.37	2.62	2.54	2.47
Mean abs. error		0.25	0.24	0.25	0.22	0.18
Mean error		-0.07	-0.04	-0.04	-0.08	-0.06

Table 5.1: The atomization energies calculated with the five different enhancement functions and the experimental values. All energies are given in eV. The different functional solutions are given as: F_x^{opt} is the local optimum to the cost function that does not violate the homogeneous gas limit for s^2, F_x^{gm} is the global optimum, and for $F_x^{(a)}$, $F_x^{(b)}$, $F_x^{(c)}$ the corresponding limit has been omitted. Table from Toftelund [2006].

[6] See Toftelund [2006]: Master project thesis with the purpose of developing a bayesian error estimation functional that would be useful for chemisorption systems.

5.4 Full machine learning approach Bayesian density functional.

Following study presented in the previous section, a more machine learning heavy approach was investigated for a Bayesian error estimation functional applicable for chemisorption.[7] In the following we will provide a concise overview of the study and the findings divided in the natural topics.

5.4.1 Datasets

To provide a better training for the fitting of the functional a number the following datasets were included

- An extended dataset of atomization energies consistent of 148 molecules from the G2/97 dataset
- A compilation of reaction energies, see table in the table to the right (figure 5.5)
- 11 chemisorption systems, with 10 of them adapted from Hammer et al. [1999]

$$\begin{aligned}
H_2 + CO_2 &\rightleftharpoons H_2O + CO \\
4\ H_2 + CO_2 &\rightleftharpoons 2\ H_2O + CH_4 \\
H_2 + CO_2 &\rightleftharpoons HCOOH \\
3\ H_2 + CO_2 &\rightleftharpoons CH_3OH + H_2O \\
3\ H_2 + CO_2 &\rightleftharpoons 3/2\ H_2O + 1/2\ CH_3CH_2OH \\
10/3\ H_2 + CO_2 &\rightleftharpoons 2\ H_2O + 1/3\ C_3H_8 \\
7/2\ H_2 + CO_2 &\rightleftharpoons 1/2\ C_2H_6 + 2\ H_2O \\
3\ H_2 + CO_2 &\rightleftharpoons 2\ H_2O + 1/2\ C_2H_4 \\
11/4\ H_2 + CO_2 &\rightleftharpoons 1/4\ \text{butadiene} + 2\ H_2O \\
2\ H_2 + CO_2 &\rightleftharpoons 1/2\ CH_3COOH + H_2O \\
2\ H_2 + CO_2 &\rightleftharpoons 1/2\ HCOOCH_3 + H_2O
\end{aligned}$$

Figure 5.5: Listing of systems in reaction energies dataset of Petzold [2010].

The addition of a dataset for respectively reaction energies and chemisorption system provides a larger coverage of the materials properties which the functional family is intended for than in the previous studies. A dataset for solid state properties was however omitted in this study.

The energy contributions of the basis functions were again based on self-consistent PBE densities.

5.4.2 Model space

The model space was, as previous studies, a parametrization of the GGA exchange enhancement factor. A linear parametrization of however suggested, as is the study of section 5.2. The parametrization was made with basis functions of the following forms

$$f_x = 1 + \kappa - \frac{\kappa}{1 + \frac{\mu s^2}{\kappa} e^{\theta \mu s^2 / \kappa}}, \tag{5.3}$$

with $\kappa = 0.804$ and $\mu = 0.2195$. The reduced density gradient s is as in equation 2.9. θ is the free parameter parametrization.

The basis functions for the model are given in figure 5.6 to the right.

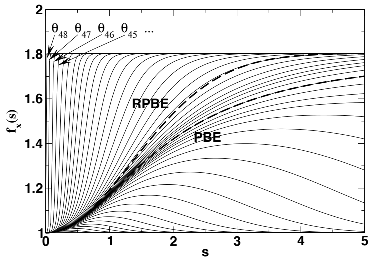


Figure 5.6: Basis functions of in Petzold [2010]. Adopted from Petzold [2010].

[7] See Petzold et al. [2012] and Petzold [2010]. This chapter will use mainly refer to Petzold et al. [2012] and occasionally to Petzold [2010] for specific details.

The linear model provides in principle a full coverage of the GGA enhancement factor model space, if enough functions are included. However, for numerical reasons only a certain number of functions can be included, and thus having a functional space that covers the relevant region of the s-range is important. In the parametrization suggested, the RPBE and PBE functional are covered like Toftelund [2006], but at the same time the model is flexible to provide enhancement factors much different from these.

5.4.3 *Model selection with regularization and bootstrap resampling*

Using a full linear parametrization with a many parameters called for ways to called for a smooth way to limit the model complexity, and thus introducing least squares ridge regression with the Tikhonov matrix for the cost function, hence RR-LS of equation 4.15 with the Tikhonov matrix of section 4.3.4.

For the Tikhonov matrix a number of different regularization priors were suggested through the formulation $R(\boldsymbol{a}, \omega_1(,\omega_2)) = \omega_1^2 P_1(\boldsymbol{a})(+\omega_2^2 P_2(\boldsymbol{a}))$ combined with different origo for the solution, see figure 5.7 for definitions and figure 5.8 for plots.

zero: $P^{\text{zero}}(a) := \sum_n a_n^2$

cPBE: $P^{\text{cPBE}}(a) := \sum_n (a_n - a_n^{\text{PBE}})^2$

dPBE: $P_1^{\text{dPBE}}(a) := \int \left(f(s;a) - f^{\text{PBE}}(s)\right)^2 \mathrm{d}s$

$P_2^{\text{dPBE}}(a) := P^{\text{zero}}(a)$

smooth: $P_1^{\text{smooth}}(a) := \int \left(f''(s;a)\right)^2 \mathrm{d}s$

$P_2^{\text{smooth}}(a) := P^{\text{zero}}(a)$

Figure 5.7: The different regularization priors tested in for functional. Adapted from Petzold [2010].

It was found beneficial to not make any explicit constraints on the functional.

The optimal regularization strength for each model was found by minimizing the Bootstrap 0.632 estimation prediction error resampling estimate, see section 4.4.3. With the Tikhonov regularization the notion of effective number of parameters was furthermore introduced as a measure of the functional complexity in the density functional fitting context, see section 4.3.1.

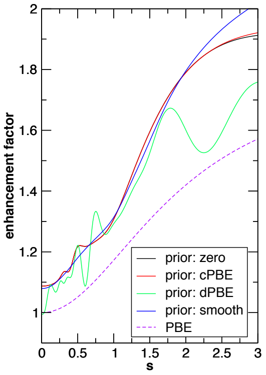


Figure 5.8: Comparing the optimal fit of different regularization priors.Adapted from Petzold [2010].

The different regularization methods were compared by fitting to the atomization energy dataset and evaluating on the atomization, chemisorption, and reaction energy datasets, see table 5.2. It was found that the smoothness Tikhonov matrix were most effective at ensuring transferability from the atomization energies to the chemisorption dataset. The transferability for the smoothness model to the reaction energies was worse than for the alternative methods. The reaction energies dataset is however very similar to the atomization energies, and the dataset is reasonably described by the smooth model, whereas the dPBE is performs very bad for the chemisorption systems. The smooth function compared to the cPBE and zero methods were overall similar average performance, but with emphasis on either chemisorption energies or reaction energies.

prior	ω^2	aMAE (eV)	csMAE (eV)	rMAE (eV)
cPBE	0.11	0.119	0.225	0.141
zero	0.11	0.119	0.224	0.137
dPBE	(0.6,0.03)	0.117	0.412	0.086
dPBE	$(1.0,10^{-7})$	0.107	9.265	0.095
smooth	(0.4,0.001)	0.121	0.176	0.188
smooth	$(0.5,10^{-7})$	0.121	0.181	0.173

Table 5.2: Comparing mean absolute error for atomization energies (eMAE), chemisorption (csMAE) and reaction energies (rMAE) datasets for different regularization priors. All functionals were fitted to the atomization energies only. Adapted from Petzold [2010].

5.4.4 *Model compromise for multiple datasets*

The inclusion of multiple datasets for fitting in the study raised the question for what compromise between the optimal model to the different datasets to make. This problem was addressed indirectly looking at how to create a better ensemble for both the G2/97 dataset and the chemisorption dataset after a fitted model was created. For two datasets, a ratio between how well the ensemble should reproduce errors of one or the other datasets were introduce as a measure of this. This provided for a handle were the ensemble that performed best to ones liking could be chosen.

5.4.5 *Discussion*

The study in Petzold et al. [2012] illustrated how the machine learning tools could be used to ensure that a transferrable functional could be created in a highly flexible model space, through the use of smoothness regularization and bootstrap resampling. The smooth fit made to the G2/97 dataset is presented in figure 5.9 along with the result of using the fit with ensemble on the ammonia synthesis. The ensemble can be seen to widen out at about s = 2-2.5, indicating that not much information is available about that region in the cost function, which in agreement with the earlier observation that the relevant s region is 0-3.

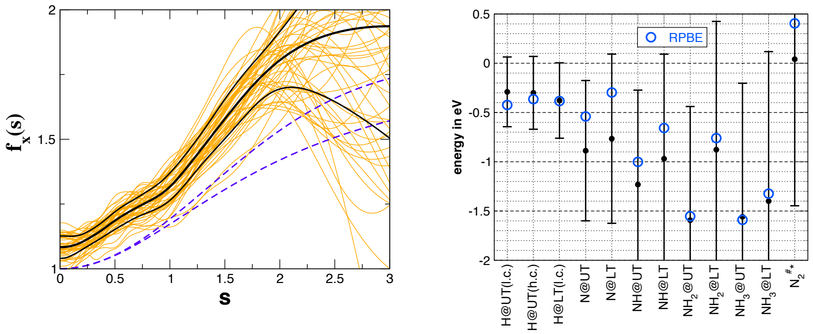


Figure 5.9: Left: The smooth fit to G2/97. The black line in the middle is the fit. the ensemble is in yellow and the one standard deviation of the ensemble is given as the surrounding black lines. Right: In black the results of the smooth fit with error bars for the different steps in Ammonia synthesis. Adopted from Petzold [2010] where more details can be found.

The inclusion of several datasets highlighted that an approach for properly treating the compromise between fitting different materials properties are called for. Going for more than two datasets will make the process of deciding a ratio between reproducing the error of different datasets

In the study the functional with ensemble was used to assess the reaction path uncertainty of ammonium synthesis. This study highlighted how that the bayesian error estimation ensemble for functionals in the future would bring valuable insights on the uncertainty in catalysis design screening studies for chemical processes.

6 The BEEF-vdW functional [1]

In the following the BEEF-vdW study will be presented, in which we set out to to use the learnings of the previous studies and create a functional competitive equally costly functionals for surface science. The fitting procedure expanded upon the datasets, the ingredients of the model, and the model selection.

For the model space, a full parametrization of the GGA space was done similar to Petzold et al. [2012], but in a basis where full orthogonally were insured, such that a very flexible model would be possible with limited numerical issues. This parametrization were made in a transformation of the reduced density gradient such that the flexibility foremost would be in the energetically relevant part of the parameter space. In addition the correlation was expanded to be a mixture of local, semi-local and non-local correlation functionals.

Several new datasets were introduced for training the model. The training data were now to include many more of the types of systems that are modeled in surface science studies, with the intention of much more deliberate ensuring that the functional were to be optimize to perform good in real use cases. This also allowed for a full benchmark against alternative functionals in a much more broad way. Furthermore, a number of validation studies were introduced, which provided further understanding of the BEEF-vdW functional and it's alternatives.

For the model selection the introduction of many datasets provided that we should take more systematically approach to the compromise between how well to describe different materials properties. This was done in a two step fitting procedure: First by finding the best fit to the individual materials properties; and in the second step, making a compromise between the fits to the individual datasets.

In the following the above workings will be described in more details, and in addition a outline for further investigation will be made, similar to the previous presentations - which will set the scene for the next study in the BEEF functional series.

[1] This chapter is based on the study of Wellendorff, J. and Lundgaard, K. T. and Møgelhøj, A. and Petzold, V. and Landis, D. D. and Nørskov, J. K. and Bligaard, T. and Jacobsen, K. W., "Density functionals for surface science: Exchange-correlation model development with Bayesian error estimation", Phys. Rev. B 85 (2012), pp. 235149.

6.1 *The parametrized GGA+vdW model space*

We employed a model space for the BEEF-vdW with a mixture of a full parametrization of the exchange enhancement factor, and for the correlation a sum of a local, semi-local and non-local functionals from the litterature. This was to insure that the exact limits of the functional could easily be enforced, while providing a much more flexible model space.

The parametrization of the enhancement factor followed the procedure of Petzold et al. [2012] for making a parametrization that would be highly flexible, and then afterwards use machine learning tools to constrain the model. The GGA exchange energy is therefore given as $E_x^{\text{GGA}}[n, \nabla n] = \int \varepsilon_x^{\text{HEG}}(n) F_x(s) n(\boldsymbol{r}) d\boldsymbol{r}$.

We decided to expand the exchange enhancement factor in a Legendre polynomial series within a transformation of the reduced density gradient. The transformation is

$$t(s) = \frac{2s^2}{q + s^2} - 1 \ , \ -1 \le t \le 1,$$

with $q = 4$ and the Legendre polynomial expansion was given as

$$F_x(s) = \sum_m a_m B_m(t(s)).$$

The Legendre polynomial basis within the transformation is illustrated in the figure to the right (figure 6.1).

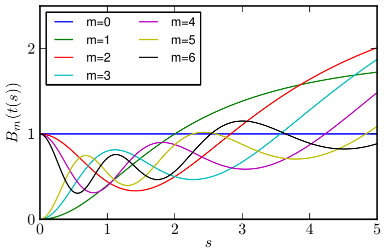


Figure 6.1: Legendre polynomial exchange basis functions $B_m(t(s))$ illustrated for polynomial orders 0 to 6. Adopted from the supplementary material of Wellendorff et al. [2012].

The transformation is a Padé approximant that was selected to almost be that of the PBE functional.[2] So that choosing $F_x(s) := 1.4 + 0.404 \cdot t(s)$ with $q := \kappa/\mu = 0.804/0.21951 \approx 3.663$ would be that of the PBE exchange enhancement factor. For this is should also be noted that the q value of PBE is not chosen to fulfill the slowly varying electron gas, but this is however the case for the PBEsol exchange functional.

The total GGA exchange energy were by the former definitions given as

$$\begin{aligned} E_x^{\text{GGA}}[n, \nabla n] &= \sum_m a_m \int \varepsilon_x^{\text{UEG}}(n) B_m(t) n(\boldsymbol{r}) d\boldsymbol{r} \\ &= \sum_m a_m \mathcal{E}_m [n, \nabla n] \,, \end{aligned}$$

where $\mathcal{E}_m$ is the exchange energy corresponding to the basis function B_m.

The parametrization of the correlation energy GGA functional space is not as simple as for the exchange, which only makes the

[2] see Perdew et al. [1996a]

parametrization in a single parameter. The parametrization of the correlation energy functional was therefore given in already established correlation functionals, hence

$$E_c\left[n, \nabla n\right] = \alpha_c E_c^{\mathrm{LDA}} + (1-\alpha_c)E_c^{\mathrm{PBE}} + E_c^{\mathrm{nl}},$$

where the LDA correlation is mixed with the PBE semi-local correlation through the parameter α_c, with the full non-local correlation of vdW-DF2. [3]

The total exchange correlation model space was therefore given in the form

$$E_{\mathrm{xc}} = \sum_{m=0}^{M_x-1} a_m \mathcal{E}_m + \alpha_c E_c^{\mathrm{LDA}} + (1-\alpha_c)E_c^{\mathrm{PBE}} + E_c^{\mathrm{nl}},$$

where $M_x = 30$, and the total number of fitting parameters is 31. Within the model space, aside from the q value, the common constraints of the GGA exchange could be invoked. This includes the uniform electron gas limit $F_x(0) = 1$ and recovery of the correct gradient expansion for slowly varying densities, and the Lieb–Oxford bound. $F_x(s \rightarrow \infty) = 1.804$ for large electron density gradients and/or small densities. The sum of LDA and PBE correlation is furthermore constrained to unity, and to fulfill $0 \leq \alpha_c \leq 1$.

6.2 *The exchange-correlation model selection*

We chose to do the fitting procedure, as before mentioned, in a two step process. Where the first is individual fits to the datasets, and the second to make a compromise between the first.

6.2.1 *Fitting individual materials properties*

The RR-LS cost function was used for the individual datasets. For the exchange enhancement factor, we employed a Tikhonov regularization with smoothness defined in the $t(s)$ transformed space. By this we would insure that the model were regularized accordantly to how we expected a reasonably exchange model to be. Ones were inserted in the diagonal of the Tikhonov matrix matrix for the zero and first order exchange enhancement polynomials and for the correlation LDA-PBE mixing parameter.

The origo of the solution for the exchange enhancement factor was $F_x(s) := 1.4 + 0.404 \cdot t(s)$ - hence very similar to that of PBE exchange. The origo for correlation was $\alpha_c = 0.75$, thus having 0.75 LDA correlation and 0.25 PBE correlation.

[3] LDA of Perdew and Wang [1992], PBE of Perdew et al. [1996a] and the vdW correlation of Lee et al. [2010].

The model complexity was found using the bootstrap .632 procedure with 500 random generated samples made individually for each regularization strength. As the regularization of the s22x5 systems for long distances gave solutions with many parameters that seemed to unphysical, we made the restriction that $F_x(s \rightarrow \infty) \geq 1$, which resulted in tuning up the regularization strength in fitting these datasets.

The expectation values were given as non-self-consistent results on RPBE densities; and RPBE geometries when geometries were not given from the reference data.

6.2.2 The individually trained exchange-correlation models

The result of the fitting procedure for the dataset used in the study are given in the table 6.3 to the right, and the exchange enhancement factor are provided in the figure 6.2 below.

The procedure is hence applied to molecular, solid state, surface chemical, and vdW dominated energetics as represented by the CE17, RE42, DBH24/08, G2/97, and Sol34Ec datasets, as well as the five S22x5 subsets.

	α_c	M_{eff}	$F_x(0)$	$F_x(\infty)$	MSD	MAD	STD
CE17	0.90	4.7	0.97	2.15	−10	96	116
RE42	1.00	4.2	1.06	1.21	19	168	207
DBH24/08	0.00	3.7	1.14	3.14	1	116	142
G2/97	0.27	7.2	1.10	2.53	−13	109	149
Sol34Ec	0.00	7.7	0.97	1.25	−4	168	208
S22x5-0.9	0.81	3.2	0.96	1.68	0	9	11
S22x5-1.0	0.82	3.1	0.98	1.87	0	8	10
S22x5-1.2	0.40	5.7	1.04	2.38	0	4	6
S22x5-1.5	0.85	4.0	1.02	1.91	−1	3	4
S22x5-2.0	1.00	3.3	0.95	1.37	2	3	3

Table 6.1: The model selection results for the individually training datasets. M_{eff} is the effective number of parameters. MSD, MAD and STD are mean signed, mean absolute, and standard deviation, respectively, all in meV. The results are non-selfconsistent.

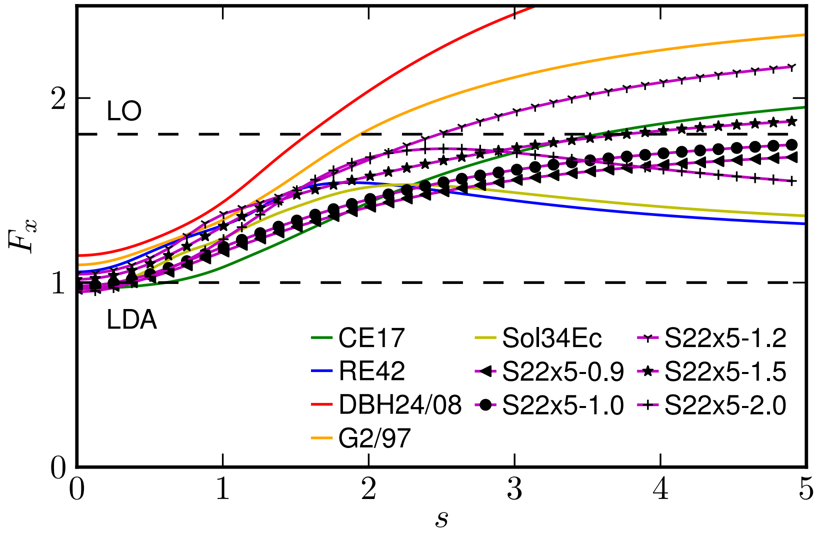


Figure 6.2: Exchange enhancement factors of the individually trained XC models listed in table 6.1.

For model to DBH24/08 it is favorable with a GGA exchange that substantially violates the LDA limit (Fx(0) = 1.14) along with inclusion of full PBE correlation ($\alpha_c = 0$). The model furthermore overshoots the Lieb-Oxford bound (LO) significantly ($F_x(\infty) = 3.14$). The exchange-correlation to the G2/97 set shows similar trends for the GGA exchange and PBE correlation, but it is less extreme.

The former are dramatically different from the Sol34Ec cohesive energies. The GGA exchange is starting from below $F_x = 1$, and then reaching a maximum at $s \approx 2$, and finally declining slowly towards $F_x = 1.25$. The model optimized for the cohesive energies uses full PBE correlation. As was previously noted, only s-values up to about 2.5 are of energetically importance, and that the GGA exchange for some of the models exceeds the LO bound for high s-values is therefore not expected to have a significant importance, however for the models where this is the case, a significant breaking of the LDA limit is also observed.

The remaining of datasets in Table6.1 has optimized exchange-correlation models that are more alike that of the common GGA functionals, with all exchange enhancement factors starting out near the LDA limit and intermediate correlation mixing parameters.

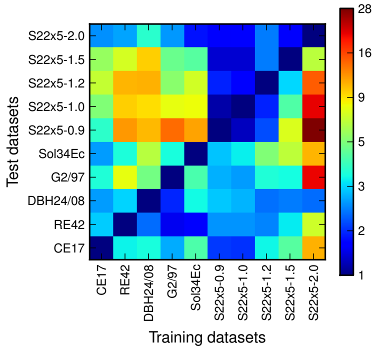


Figure 6.3: The relative standard deviations obtained when non-self-consistently applying the exchange-correlation models found individually for each training dataset, listed on the abscissa, to all 10 training datasets along the ordinate.

To test the transferability of the individual models for the datasets all models for the individual training were applied on all other training sets. For comparing the transferability we reported the relative root mean square deviation, denoted rSTD for relative standard deviation, for all individually optimized models to all the datasets, see figure to the right.

The rSTD is thus a measure of model transferability and the diagonal in the figure, from bottom left to top right, is, by definition, ones. The map features two distinct areas of mostly reddish squares: To the far right, the S22x5- 2.0 model yields rSTD > 5 for all other sets than DBH24/08, and rSTD 28 for S22x5-0.9. Furthermore, a 5×4 square in the top left corner illustrates that exchange-correlation models trained on chemical or solid state datasets perform significantly worse on vdW type energetics, than models fitted to the latter. The S22x5-2.0 rSTDs are largely unaffected by changing exchange-correlation models (top horizontal row). We propose that the small density–density overlap between many of the S22x5-2.0 complexes combined, means that the nonlocal correlation is most important for these systems, and in all 10 models the none-local correlation is the same.

The individually training dataset optimized models, fitted with the method described earlier, is capable of providing models that shows much better statistics than functionals of the same model complexity provided in the litterature.[4] However it is found that the transferability of the optimized functionals between the training datasets are in many cases bad, see figure 6.3, this is however especially observed to and from the low density-density overlapping systems of S22x5.

[4] For comparison see the statistics of the benchmark provided in the following.

6.3 *Compromise model for several materials properties*

In the BEEF-vdW study it was chosen to make a model compromise using the cost functions for the individual datasets, given in equation 4.51. The optimal model would be one where decreasing the cost on one dataset would increase the relative cost on the other datasets commutative more. The model compromise is therefore that of the product of the individual costs of the datasets, provided as the first model compromise in the machine learning chapter.

The compromise between the cost of two datasets for the optimal model using the product of the costs procedure is illustrated in the figure 6.4 to the right.

The plot shows how optimizing for one property will severely increase the relative cost on the other property, where the properties are quantified by the fit to a dataset. For the product of the cost minimum it is observed that a the fraction between the two properties in a cost function that is the sum of the two cost functions is relatively flat around the optimum.

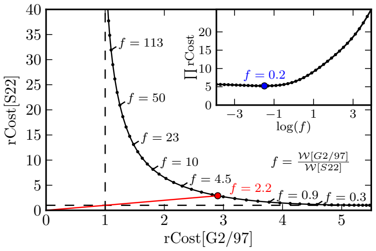


Figure 6.4: Main panel: The exchange-correlatoin model compromises between the G2/97 and S22x5-1.0 datasets illustrated in terms of relative costs (rCost) for both sets when the weight fraction f = $\mathcal{W}[G2/97]/\mathcal{W}[S22x5-1.0]$ is varied and the summed cost function is minimized where the weight is given by the fraction. A red dot marks the point of equal rCost. The fact that an XC model with $rCost[G2/97] = rCost[S22x5-1.0] = 1$ is not obtainable which illustrates the necessity of a model compromise. Insert: The product of relative costs display a minimum (blue dot) for a certain weight fraction.

6.4 *The BEEF-vdW functional*

The BEEF-vdW functionals was designed with the weights on the individual datasets given in table 6.2. The weights were modified to take into account for how important the different materials properties are and how much different datasets were describing the same property. The RE42 set is based on G2/97 molecules, and the data in RE42 is therefore correlated with some of the data in G2/97, and it was therefore decided to weight those two datasets by each one half. Similarly for the S22x5 subsets, where the same complexes are found in all 5 datasets. The weight was set to half of the natural weight for these datasets, $1/5 = 0.2$, as it was observed that this would benefit the functional towards being better for surface chemistry.

	w	w/C	rCost	rSTD
CE17	1.0	1.80	1.7	1.3
RE42	0.5	0.62	2.5	1.8
DBH24/08	1.0	0.65	4.9	2.3
G2/97	0.5	0.62	2.6	1.6
Sol34Ec	1.0	0.43	7.5	2.8
S22x5-0.9	0.1	0.01	28.6	5.4
S22x5-1.0	0.1	0.04	9.1	2.9
S22x5-1.2	0.1	0.09	3.5	2.1
S22x5-1.5	0.1	0.08	4.1	2.1
S22x5-2.0	0.1	0.18	1.8	1.5

Table 6.2: The BEEF-vdW model compromise overview. The effective weight in determining the exchange-correlation model solution is w/C for each dataset, as iteratively found from minimizing the product cost function. The relative standard deviation (rSTD) is the ratio of the STD at the BEEF-vdW compromise to the STD at the regularized individual solution in Table 6.1. The relative cost (rCost) is again relative to the individual fit where the the regularization strength of the individual fit is used.

The parameters to the resulting functional are provided in table 6.2, showing the w/C, rCost, and rSTD for all the datasets used in training the model. The is observed that the model compromise model is significantly worse for the S22x5-0.9 dataset than the individually trained model, which is both made visible in the relative cost and in the relative STD. The rest of the S22x5 set is much more adaptable to the compromise model, and on the same level of relative STD as the solids dataset and the barriers.

The expansion coefficients for the BEEF-vdW functional is given in table 6.3. It is here seen that the smoothness regularization as expected suppresses the higher order polynomials. The correlation consist of 0.6 LDA and 0.4 PBE, and the non-local correlation of vdW-DF2.

m	a_m	m	a_m
0	1.516501714	15	$-8.018718848\times10^{-4}$
1	4.413532099×10^{-1}	16	$-6.688078723\times10^{-4}$
2	$-9.182135241\times10^{-2}$	17	1.030936331×10^{-3}
3	$-2.352754331\times10^{-2}$	18	$-3.673838660\times10^{-4}$
4	3.418828455×10^{-2}	19	$-4.213635394\times10^{-4}$
5	2.411870076×10^{-3}	20	5.761607992×10^{-4}
6	$-1.416381352\times10^{-2}$	21	$-8.346503735\times10^{-5}$
7	6.975895581×10^{-4}	22	$-4.458447585\times10^{-4}$
8	9.859205137×10^{-3}	23	4.601290092×10^{-4}
9	$-6.737855051\times10^{-3}$	24	$-5.231775398\times10^{-6}$
10	$-1.573330824\times10^{-3}$	25	$-4.239570471\times10^{-4}$
11	5.036146253×10^{-3}	26	3.750190679×10^{-4}
12	$-2.569472453\times10^{-3}$	27	2.114938125×10^{-5}
13	$-9.874953976\times10^{-4}$	28	$-1.904911565\times10^{-4}$
14	2.033722895×10^{-3}	29	7.384362421×10^{-5}

Table 6.3: Expansion coefficients a_m for the BEEF-vdW Legendre exchange basis functions B_m. The correlation mixing parameter, α_c was 0.6001664769.

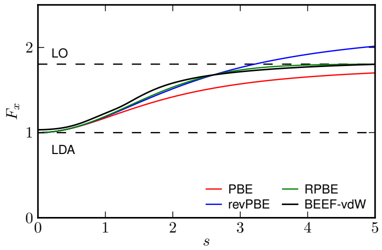


Figure 6.5: The BEEF-vdW exchange enhancement factor compared to those of a few standard GGA exchange functionals. The corresponding BEEF-vdW correlation functional is composed of 0.6 LDA, 0.4 PBE, and the vdW-DF2 nonlocal correlation.

In figure 6.5 we see the the exchange enhancement factor plotted with a couple of GGA functionals that we find relevant comparisons. It is observed how the shape is very different in that does not fulfill the LDA limit, but follows the steepest functionals up for thereafter to level off. The limits for $s = 0$ and $s = \infty$ are 1.034 and 1.870, respectively.

6.5 *The Bayesian error estimate*

A Bayesian error estimating ensemble of functionals were created, using the formalism of section 4.3.2. The ensemble yet again confirms that the loss function only provide guidance for the functional for s up to around 2.5, after this the ensemble spreads out, and it is only the regularization that is defining the behavior. It can also be observed that the uncertainty of the functional around the LDA limit includes for one standard deviation the LDA limit. The correlation is within one standard deviation within the bounds of $0 \leq \alpha_c \leq 1$.

In table 6.4 we compare the resulting error estimations of the training datasets with the actual errors for the datasets when the BEEF-vdW functional is applied self-consistently. The estimates are of varying quality, however for all of them they provide a order of magnitude error for the materials property error. The quality of the BEEF-vdW error ensemble estimations can be put in perspective to that no other systematic error estimates are currently available to our knowledge. In Wellendorff et al. [2012] we illustrates how one can use calculations of different functionals to estimate the error. However it was observed that the ensemble provided a comparable view of the actual errors, but with the benefit of being much computational cheaper and systematic.

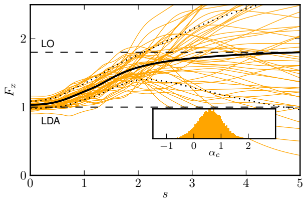


Figure 6.6: The bayesian ensemble of the exchange-correlation functionals around BEEF-vdW. In the main panel the Black solid line is the BEEF-vdW exchange enhancement factor, while the orange lines depict $F_x(s)$ for 50 samples of the randomly generated ensemble. Dashed black lines mark the exchange model perturbations that yield DFT results ±1 standard deviation away from BEEF-vdW results. The insert provides a histogram of the distribution of correlation parameters in an ensemble containing 20,000 samples. The distribution is centered around $\alpha_c = 0.6$.

	BEEF-vdW	Ensemble estimate
CE17	143	209
RE42	372	253
DBH24	331	144
G2/97	242	312
SolEc34	576	436
s22x5-0.9	171	197
s22x5-1.0	94	181
s22x5-1.2	36	137
s22x5-1.5	8	67
s22x5-2.0	5	18

Table 6.4: A comparison of the self-consistent BEEF-vdW standard deviations to those predicted by the Bayesian error estimating ensemble of BEEF-vdW. All energies in meV.

6.6 *Assessment of BEEF-vdW functional*

An extensive benchmark overview was created, where the BEEF-vdW functional was compared to other commonly used semi-local functionals and the hybrid functional B3LYP. Furthermore several qualitative studies were provided to show how the functional performed in relevant surface science studies, where higher level reference data of other functionals are available. In this thesis however, the main point is to present the fitting procedures for exchange-correlation functionals and provide a throughout assessment of those. Therefore only a single assessment plot will be given her, as we believe that it is representative for the functional within this context, and further benchmark for the includes BEEF-vdW in the following chapters.

In figure 6.7 a barplot with the Mean absolute error for the benchmark datasets with self-consistent BEEF-vdW calculation results, as well as results for the functionals that we in the study found most relevant to compare to.[5]

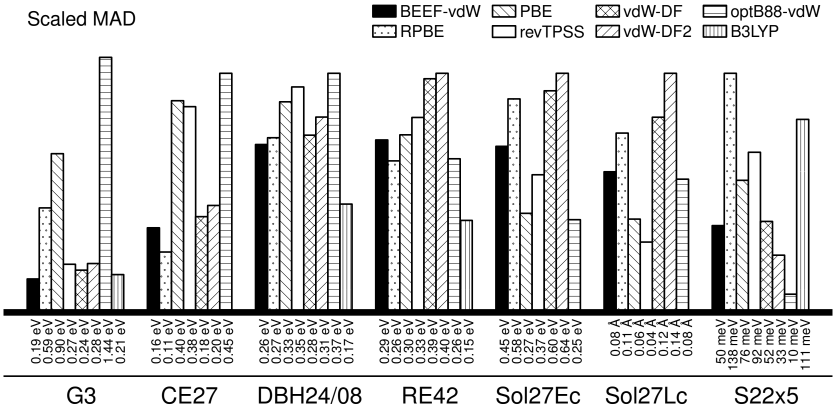


Figure 6.7: Bar plot comparison of the accuracy of different density functionals in predicting various materials properties. For each dataset, the bars illustrate proportionally scaled mean absolute deviations. B3LYP calculations were not performed for extended systems.

The overview illustrates that the BEEF-vdW overall performs well compared to current functionals. The functionals given by type are GGA: PBE and RPBE; revTPSS is a meta-GGA; vdW-DF, vdW-DF2 and optB88-vdW are of GGA+vdW; and B3LYP is a hybrid.[6]

A number of the datasets have been expanded in the benchmark, and the reason for not using them as training sets were due to them not being available when the BEEF-vdW fit were made.[7]

The performance comparison to the other functionals can be observed to match how the weighting to the different training sets were done. A good performance is achieved on G3, where the performance outliers comes from that a subset of the G3 are of very large systems

[5] See calculation details in Wellendorff et al. [2012].

[6] PBE in Perdew et al. [1996a], RPBE in Hammer et al. [1999]; revTPSS in Perdew et al. [2009]; vdW-DF, vdW-DF2 and optB88-vdW in Dion et al. [2004], Lee et al. [2010], Klimes et al. [2010]; B3LYP in Stephens et al. [1994].

[7] See chapter 3 for more details on the datasets.

that can skew the statistics. It is noted that the BEEF-vdW functional was fitted to the G2/97 dataset, but it was found that this was very transferable to the atomization energies of the G3 dataset. The BEEF-vdW performs almost equal to the RPBE for CE27 and for the RE42, which was hoped for. The performance in s22x5 on the level of vdW-DF but it not as good as the OptB88-vdW functional that has been specifically optimized for this dataset. The solid state performance is given in by the Sol27Ec (cohesive energies), and the lattice parameters of Sol27LC. For these it is seen that the BEEF-vdW performs significantly worse than the PBE and revTPSS, that both to a much higher degree fulfills the slowly varying electron gas limit.

6.7 *Summary and discussion*

The BEEF-vdW functional study illustrated how to bring earlier insights together fit a general applicable functional.

6.7.1 *The strengths of the BEEF-vdW study*

What foremost made the BEEF-vdW study different from the former Bayesian error estimating functional studies was the number of materials properties that was covered in the training sets. For the first time we had a full coverage of most of the relevant surface science properties in the training datasets. This allowed use to see how the different properties were to be describe in the model space of GGA exchange with GGA+vdW non-local, and further how a compromise could come together for simultaneously treating all these properties at the same time.

For the fitting the individual datasets, the machine learning tools that was presented in Petzold et al. [2012] were fully put in use. This included Tikhonov regularization with a smooth basis, and with the model complexity given by minimizing the Bootstrap 0.632 estimated predicted error resampling estimator. However, with a transformation of the s parameter with a Padé approximant the smoothness was defined in a way that matches the expectation for the final functional. Furthermore, the usage of the a Legendre polynomial basis within the transformed s parameter space, ensured would be able to reach a higher complexity that previous, if needed. And, the correlation of a mixture between the LDA and PBE correlation illustrated a simple way to form a correlation functional to match the vdW non-local functional to exceed the performance of it.

The model compromise of this study, for the first time to our knowledge, illustrated how a more systematic and transparent view on the weighting of the different materials properties could be done for the fitting of an exchange-correlation functional; where the weights of the different materials properties were not to be set by intuition of the creators, but by explicitly stating how much we wish to optimize for individual properties through the product cost weights.

Again it was clear that not fulfilling the LDA limit makes sense for these fitted functionals, as the performance penalty would be very big if that constraint were to be restricted.

Lastly the study also provided a full benchmark for the functional to easily enable a comparison to other relevant functionals.

6.7.2 *Limitations of the BEEF-vdW study*

There were however also a number of issues number of limitations in the study that should be highlighted.

In the study the datasets used in the benchmark of the different functionals were largely extended compared to the training datasets. This made it possible to validate the transferability of the optimal functional. However it would have been better if the extra data were to be used in training the data, and then having the cross-validation algorithm ensure that the model were to be constrained throughly.

It also becomes apparent that fitting the G2 or G3 datasets and evaluate on this dataset is not a good validation of the overall performance of a fit. In the G3 datasets (concealing the G2) a number of large carbon chains makes functionals such as the OptB88-vdw come out very bad as it has not been optimized for these systems. However this functional is able to capture the relevant molecular reaction chemistry through the RE42 dataset, which is what is relevant for surface science studies.

For the GGA-vdW model space it was also observed that the barriers could not be properly described in the model space. It is known that the hybridization happening in molecular transition states inflicts a high degree of self-interaction if not a self-interaction free functional is used. Furthermore, it was observed that the optimized functional for this dataset were very different from the optimal models to the other datasets. It was most comparable to the G2 optimal fit, which following the discussion from the previous paragraph is not in high standing either. It could therefore be argued that in the current form of the model compromise, such a dataset should be omitted.

The performance of the functional was heavily limited by the exchange model space, which was also known from the previous studies of the Bayesian error estimating functionals. The improvements in performance that was gained overall, compared to the functionals that were benchmarked against, we attribute to a different optimization criteria combined with not fulfilling the LDA limit of the exchange enhancement factor and having a mixed correlation of LDA and GGA. It is well understood that the meta-GGA functional space would make it possible to create a higher performing functional, with only a modest gain in computational overhead.

A more flexible correlation parametrization would also be preferred. Holding the sum of the LDA and GGA correlation equal to one does not make us fulfill the LDA limit as the exchange factor is not constrained in the limit, so this limitation of the parametrization could harm the performance of the optimized functional, while not providing the theoretical justification that was intended by making it.

For the model selection, the following limitations should be mentioned. For the individual fits it was observed that for several datasets, the Bootstrap procedure did not provide an optimal model complexity. However, following the study it was found that there was an error in the implementation of the ERR in the bootstrap EPE estimation, where the loss function was that of the variance, and thus the function did not take the bias into account. It is not expected that this problem had a significant negative effect on the BEEF-vdW functional solution, as the additional constrain of $F_x(\infty) > 1$ were added to S22x5, and because of the final model compromise. In the study new samples were made for every regularization strength, which is not optimal as explained in the machine learning chapter. However this is not expected to have influenced the model complexity optimum, as it was countered by using many bootstrap samples, and by careful oversight.

The use of the model compromise in the form of the product of the cost functions of the individual datasets was that of a compromise, and it did not use the full information at hand to provide a maximum likelihood solution for the full problem. The compromise did yield a well performing functional, and for this model space. If one datasets's individual fits would have had a very high regularization associated with it, then it would have made the compromise model much more constrained too.

A last limitation, that was not addressed in the BEEF-vdW model fit, was the effect of self-consistent of the fit. It was observed that the root mean square deviation on the sum of S22x5 increase by about

50% from the prediction of the fit to the self-consistent results. In Petzold [2010] it was argued that fitting on the results of the current BEEF-vdW densities would have lead to close to a convergence between the non-self-consistent results and the self-consistent results. A further investigation of this would have added clarity to this issued, and insured that the final fit would indeed be the optimal fit in terms of self-consistent model predictions.

7 A meta-GGA Bayesian Error Estimation Functional

We will now take a look at the next functional of the BEEF class, called mBEEF. Most notably this functional is optimized in a parametrization of the Meta-GGA model space, and the model compromise is that of the geometric mean loss function of the datasets. These extensions were made to overcome many of the limitations that were observed in the development of the BEEF-vdW functional.

In this chapter we look into the details on the methodology of the mBEEF functional, and see how it is benchmarked against other functionals, similar to the BEEF-vdW study. Furthermore we will take a look at a qualitative study of the mBEEF functional, where it can be observed how the different rungs of Jacobs Ladder dictate what accuracy can be achievable within the normal formalisms.

7.1 The model space

For the mBEEF functional we created a full parametrization of meta-GGA (MGGA) exchange energy. We define the model space following the definitions given for the MGGA exchange rung on Jacobs ladder in the DFT chapter.

For the MGGA exchange factor we therefore write that $F_x(n, \nabla n, \tau) = F_x(s, \alpha)$, and the model space is expanded in P Legendre polynomials B depending on s and α by the transformations t_s and t_α given here:

$$\begin{aligned}
t_s(s) &= \frac{2s^2}{q+s^2} - 1, \\
t_\alpha(\alpha) &= \frac{(1-\alpha^2)^3}{1+\alpha^3+\alpha^6}, \\
P_{mn} &= B_m(t_s)B_n(t_\alpha), \\
F_x(s,\alpha) &= \sum_{m=0}^{M}\sum_{n=0}^{M} a_{mn}P_{mn},
\end{aligned}$$

where $M = 7$ and we therefore have $M_{tot} = (M+1)^2 = 64$ exchange basis functions with expansion coefficients a_{mn}. Both t_s and t_α are confined to the interval $[-1, +1]$, such that the Legendre polynomial expansion is fully orthogonal. We choose $q = \kappa/\mu = 0.804/(10/81) = 6.5124$ in the t_s transformation, and it is therefore given as a Padé approximant to the PBEsol exchange enhancement factor. The PBEsol exchange transformation was chosen for the s parameter because it makes it easy to reproduce a functional with the exact constraint of the LDA limit, and for the slowly varying electron gas gradient expansion around $s = 0$. The transformation t_α is likewise chosen so that the second order gradient expansion could be fulfilled.[1] The full exchange energy expansion is of the form

$$\begin{aligned} E_{xc} &= \sum_{m,n}^{M} a_{mn} E_x^{mn} + E_c^{\mathrm{GGA}}, \\ &= \boldsymbol{x}\boldsymbol{a}^T + E_c^{\mathrm{GGA}}, \end{aligned}$$

where $\boldsymbol{x}$ is a vector of the exchange basis function energy contributions. The correlation energy functional E_c^{GGA} is that of PBEsol.

7.2 Training datasets

In this study the number of datasets were limited to five, for which some had been updated or slightly modified since the BEEF-vdW study.

The datasets are the G3/99 molecular formation energies, and the related RE42 reaction energies to represent gas-phase chemistry. The two datasets have however in this study been modified to normalize the data better in the following way: The G3/99 were standardized with the factor $1/(N_a - 1)$, where N_a is the number of atoms in each molecule. The RE42 were similarly standardized with the factor $1/(N_r - N_p + 1)$, where N_r and N_p are the number of reactants and products in each reaction, respectively.

Surface chemistry were represented by the CE27a chemisorption energies of simple adsorbates on late transition metal surfaces. The CE27a is the CE27 dataset from earlier, where the reference is to the free atoms rather than the gas-phase adsorbates.

The functional was trained on the Sol54Ec dataset, and on the derivatives cohesive energies, with respect to the crystal volumes around equilibrium, with the experimental lattice constants from the Sol58LC dataset. The Pb data point was however excluded from the training sets however, as it was found to be a heavy outlier.

The structural geometries[2] and electronic densities for all the datasets were defined by self-consistent PBEsol functional calculations.

[1] See Jianwei Sun, Bing Xiao, and Adrienn Ruzsinszky. Communication: Effect of the orbital-overlap dependence in the meta generalized gradient approximation. *The Journal of chemical physics*, 137 (5):051101, August 2012b. ISSN 1089-7690. DOI: 10.1063/1.4742312. URL http://www.ncbi.nlm.nih.gov/pubmed/22894323

[2] This only applied for the Chemisorption systems where the structures were not given directly from the reference data.

7.3 Exchange model selection

To optimize the model, we used the the loss function of the geometric mean loss function of the least squares for the individual training datasets, given in equation 4.56. The weights on the individual datasets were all set to 1 in the loss function. For the prior solution we employed the Tikhonov transformation to the problem with the smoothness of the 2-dimensional transformed space of t_s and t_α. The Tikhonov matrix $\mathbf{\Gamma^2}$ was therefore defined by the Laplacian $\widetilde{\nabla}^2$ of the exchange basis functions $P(t_s, t_\alpha)$,

$$\begin{aligned} \widetilde{\nabla}^2 &= \frac{\partial^2}{\partial t_s^2} + \lambda \frac{\partial^2}{\partial t_\alpha^2}, \\ \mathbf{\Gamma}^2_{mnkl} &= \int_{-1}^{1} \int_{-1}^{1} dt_s \, dt_\alpha \widetilde{\nabla}^2 P_{mn} \widetilde{\nabla}^2 P_{kl}, \end{aligned}$$

where $\lambda = 10^2$ scales the regularization penalty between polynomials in t_s and t_α.

The origo of the solution was chosen to that of $F_x(s, 1) = 1$ for all s and half of the MS0 exchange along the $F_x(0, \alpha)$ model space direction. This origo is significantly different from what has been used in the BEEF-vdW study, but it was found that the ensemble behaved well in the limit of almost no information in the origo, in contrast to when the origo of MS0 or similar, where the ensemble would very constrained in much of the energetically constraint region for then to open much more out outside of this region.

The optimal model complexity was found using the cluster-LOOCV for the five datasets with the sum loss of the samples, given in equation 4.61.

7.4 *The mBEEF functional*

In figure 7.1 to the right, we show the regularization curve for sum loss estimated prediction error using the cluster-LOOCV method of equation 4.61. The model complexity is chosen as the minimum to the regularization curve. The minimum is seen to be well defined but with a two other almost as good solution for a larger model complexity. The optimal model complexity for the mBEEF model has 8.8 effective parameters

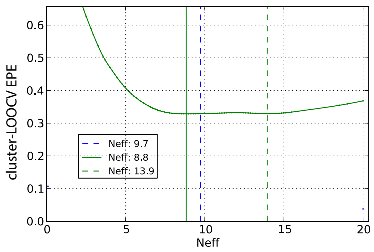


Figure 7.1: The cluster-LOOCV estimated prediction error as a function of the model complexity. The sum of the errors is used in the prediction error estimate.

In figure 7.2 the mBEEF functional is plotted along with the solution to the cost function for different regularization strengths; resulting in a range of different exchange model complexities. For the mBEEF model we observe that it is also preferable for a functional in the MGGA model space to break the LDA limit slightly.

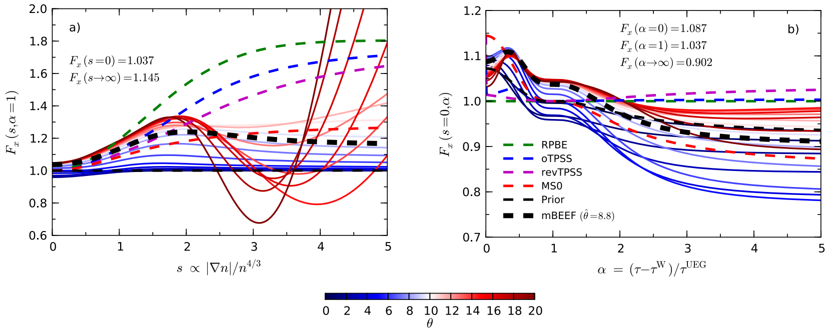


Figure 7.2: Model-compromise optimized mBEEF type exchange enhancement factors for increasing number of effective parameters θ. Full black lines indicate the chosen mBEEF $F_x(s,\alpha)$. Standard GGA and MGGA exchange functionals are indicated by dashed lines along with the prior model (black dashes). a) Projections along s for $\alpha = 1$. b) Projections along α for $s = 0$. The effective number of exchange model parameters range from 0 (dark blue) to 20 (dark red). The mBEEF model has 8.8 effective parameters.

The ensemble for the mBEEF is plotted in figure 7.3, and it takes a similar shape to that of the BEEF-vdW in the $\alpha = 1$ cut, and it is illustrated that much of the functional is mostly constrained in the low α limit, which is the region between the homogeneous electron gas limit ($\alpha = 1$) and the single atomic orbital limit ($\alpha = 0$).

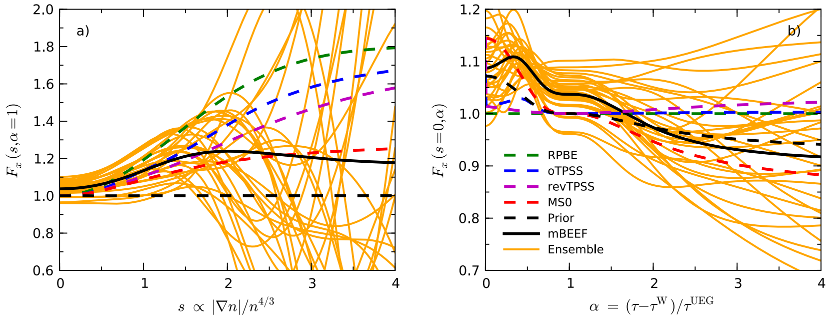


Figure 7.3: Bayesian ensemble of exchange models (yellow) around the mBEEF (solid black). a) Projections along s for $\alpha = 1$. b) Projections along α for $s = 0$.

7.5 Benchmark

In figure 7.4 we summarize a broad benchmark of popular or recent GGA and MGGA density functionals, and also mBEEF and BEEF-vdW. The bars indicate the logarithms of scaled mean-absolute errors on the five datasets applied in mBEEF training. The mBEEF exchange model compromise appears excellent: The MAE is among the three lowest for all five properties and considerably improves over BEEF-vdW in predicting lattice constants of bulk solids and their cohesive energies while retaining a good description of the adsorbate–surface bond strengths in CE27a.

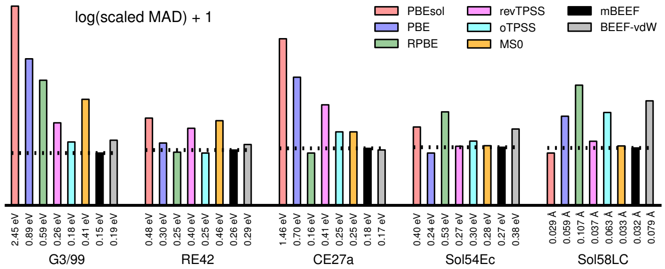


Figure 7.4: Benchmark of mBEEF against popular or recent GGA and MGGA density functionals in terms of mean-absolute errors (MAE) on predicting the materials properties represented by the 5 datasets applied in mBEEF training. BEEF-vdW is also included. Note that each bar is normalized with the smallest one and plotted on a logarithmic scale for reasons of clarity. Horizontal black dash-dotted lines indicate the mBEEF level, which is among the 3 lowest for all 5 materials properties.

7.6 The model compromise investigated

In the BEEF-vdW study, it was highlighted how a model compromise between different material properties had to be made, as the GGA model is limited. However, with the added ingredient for the MGGA functionals this compromise can be broken, and a functional that outperforms GGA functional on several material properties is possible. We further illustrated this in the study of mBEEF.

In figure 7.2 a broad selection of GGAs, MGGAs, and vdW-DF type functionals are applied in calculations of four different quantities; chemisorption energies of small molecules on close-packed transition metal facets, surface energies of various facets, solid bulk moduli, and gas-phase reaction energies. These materials properties are represented by the CE27a, SE30, BM32, and RE42 datasets.[3] The tested GGAs are PBEsol, PBE, and RPBE, while the literature MGGAs are TPSS revTPSS oTPSS and MS0. The three chosen van der Waals functionals vdW-DF, optPBE-vdW, and C09-vdW are equivalent except for the choice of PBE-like exchange. Figure 7.2a plots root-mean-squared prediction errors (RMSEs) on chemisorption energies against those on surface energies. The points within each class of exchange-correlation model space fall approximately on straight lines, illus-

[3] The SE30 and BM32 are from G. H. Jóhannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen, and J. K. Nørskov. Combined electronic structure and evolutionary search approach to materials design. *Phys. Rev. Lett.*, 88(25):255506, 2002

trating the compromise between accurate adsorbate–surface bond strengths and surface stabilities. However, the MGGA model space offers the most attractive compromises; the green line in figure 7.2a is by far closest to the origo. This is supported by figure 7.2b, where mean-signed errors (MSEs) on predicted bulk moduli are plotted against those on molecular reaction energies. The relations between mean errors are again approximately linear and the MGGA points fall closest to origo, though not all on the same straight line.

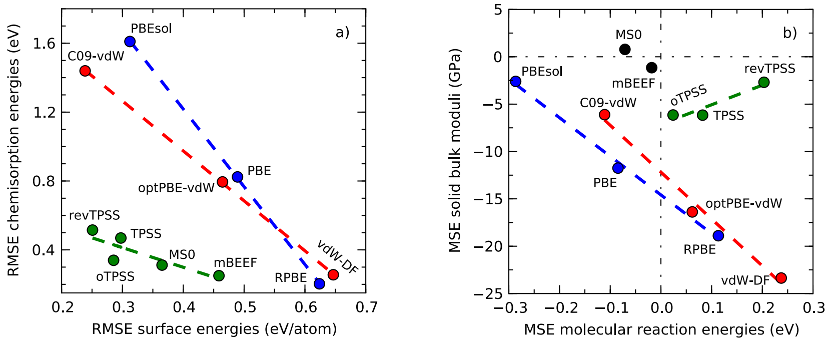


Figure 7.5: Bivariate analyses of semi-local DFT prediction errors for chemical and bulk materials properties. a) Root-mean-squared error on the CE27a chemisorption energies against that on the SE30 surface energies. b) Mean-signed error on the BM32 solid bulk moduli against that on the RE42 gas-phase reaction energies. Straight lines are fits through the GGA (blue), meta-GGA (green), and vdW-DF type (red) data points.

The bivariate prediction error analyses in figure 7.2 exceedingly confirm conjectures from other studies. The linear relationships between DFT-predicted chemisorption energies and surface energies have appeared in recent literature for the particular cases of CO on Pt(111) and Rh(111).[4,5,6] However, bivariate analyses of DFT prediction errors for surface chemistry and stability has, to the authors' knowledge, not previously been considered on such firm statistical footing as in figure 7.2.

7.7 *Summary and discussion*

We summarize the improvements of the mBEEF study compared to the BEEF-vdW study her, and provide again guidance towards further improvements of the BEE development methodology.

The benchmark of the mBEEF functional, presented a very convenient case for the mBEEF functional. The functional was amount the best performing functionals on all the datasets simultaneously, and the best performing for three of the five datasets. Several benchmark and qualitative validation studies were made in addition to the ones presented in this chapter. These overall showed that the mBEEF functional performed in comparable to other high-performing MGGA functionals, which validates that the mBEEF functionals's performance is transferable, and not an overfit to

[4] A. Stroppa and G. Kresse. The shortcomings of semi-local and hybrid functionals: what we can learn from surface science studies. *New J. Phys.*, 10 (6):063020, 2008

[5] J. Sun, M. Marsman, A. Ruszinszky, G. Kresse, and J. P. Perdew. Improved lattice constants, surface energies, and CO desorption energies from a semilocal density functional. *Phys. Rev. B*, 83:121410(R), 2011

[6] L Schimka, J Harl, a Stroppa, a Grüneis, M Marsman, F Mittendorfer, and G Kresse. Accurate surface and adsorption energies from many-body perturbation theory. *Nature materials*, 9 (9):741–4, September 2010. ISSN 1476-1122. DOI: 10.1038/nmat2806. URL http://www.ncbi.nlm.nih.gov/pubmed/20657589

the training datasets, that also constituted the benchmark sets.

In the model compromise study of bivariate prediction errors we presented a more statistically founded view of the model compromise of the different rungs of Jacobs ladder, where was illustrated how each model class are bound to make a compromise between how well it predicts high accuracy reference energetics of various material properties.

7.7.1 Strengths

The mBEEF study illustrated how the full parametrization could be made of the MGGA model space as a natural progression of the parametrization with BEEF-vdW. Where the inputs to the were transformed, so that they range would match that of the Legendre polynomial basis, and next use a 2-dimensional Legendre polynomial basis to expand this transformed 2-dimensional space. This expansion allowed for an easy definition of also the smoothness of the prior with one parameter to scale the regularization strength ratio between the s and α model spaces.

The training datasets used for the model optimization were only modestly expanded or improved to those of the benchmark of BEEF-vdW, but they now cover the range of simple energetic materials properties that are viewed as important for surface science.

The model compromise was that of the geometric mean with the least squares loss for the individual datasets, with a prior of the smoothness Tikhonov matrix in the 2-dimensional model space. The cost function of the fitting problem therefore allowed for the full data information of all the datasets to be taken into account. With an optimal model complexity found by the minimizing for the cluster-LOOCV estimated prediction error it was possible to locate a model complexity with good fitting performance but without seemingly overfitting.

7.7.2 Limitations

For the model space, it was decided to simplify the correlation functional energy to be that of the PBEsol correlation, and only the effect of the exchange were therefore given. This made it convenient for analyzing the fitting procedure, as the regularization was only to be done in the exchange enhancement space, but it did not provide the full model complexity at the MGGA rung. The exchange space of the mBEEF could cover some of the correlation, but the exchange model does not have the full information of the spin density to allow for full a correlation model as those of other MGGA functionals. Furthermore, the exclusion of the vdW non-local correlation makes the functional unequipped for studies where vdW dispersion is important.

For the training datasets, it is noted that the mBEEF functional was notably not fitted for the the S22x5 van der Waals complexes, which goes back to the discussion above, hence there were no non-local correlation in the functional form of mBEEF, so it was not expected that the functional would be able to capture the vdW interaction. Another observation in the study was however that the Pb data point in solids training dataset was an outlier. The data point was therefore excluded from the training dataset manually. If data points are outliers in the dataset because e.g. there is a large experimental error to them, then they will deter the optimization algorithm through the resampling algorithm, and we will not end up with the optimal transferable functional. The approach that was taken in the mBEEF study of manually removing an outlier is not very unsystematic, with masking it is hard to detect these outliers, see discussion in section 4.6. We would like to have methods that systematically can prevent outliers from deter the model, and remove masking effects.

The optimal model complexity was found as the global minimum to the estimated prediction error for different regularization strengths, presented in the regularization curve. However, several more complex models local minima were given for higher model complexities. This was unsatisfying as it did not make it clear if the model complexity selected is indeed the optimal, as the regularization curve could possible tilt with small changes in the dataset or in the prior.

8 Robust fitting of exchange correlation functionals

In this final study we have extended the machine learning tools to deal with some of the issues that persisted in the former studies of BEEF-vdW and mBEEF. These tools will be presented and used to fit a functional of the mBEEF-vdW, which is the natural step forward, but with the emphasis on the evaluation of the machine learning tools. The study is a work in progress.

We will in this study introduce a loss function for handling outliers, and furthermore propose a scheme for which to incorporate with the methods used in the BEEF functionals. The overall loss function will therefore still be of the geometric mean to the individual datasets of BEEF-vdW and mBEEF, but where each dataset are now represented by a robust loss function.

For the loss function, we will also propose a scaling of the data points internally in datasets that is based on the covariance matrix of our Bayesian error estimation method.

In addition we will take use of the hierarchical bootstrapping estimate with the geometric loss function, so that cost function of the optimal model correspond to the model that is optimized for in the cross validation estimating prediction error. The outlier detection is tested on a dataset with artificial outliers; to test the procedures in a controlled environment.

We will further provide detailed view of the functional models performance on the training dataset to assess the loss functions. And propose using a cross-validation estimate of the transferability of the functional to judge which loss function performs best. We will in the end propose the functional form of the mBEEF-vdW functional and make a non-self-consistent benchmark of it.

8.1 The model space

The model space for this study was for the exchange identically to that of the mBEEF study. Thus for the MGGA exchange basis we use a basis of $8 \times 8 = 64$ Legendre polynomials in the t_s and t_α

transformations.

We have tested the larger basis of 10×10 Legendre basis functions, but we did not find any significant performance improvement for the optimal model, compared to a 8×8 basis. Having a larger basis with higher order Legendre polynomials will however come at a computational cost.

The correlation was parametrized as

$$E_c\left[n, \nabla n\right] = \alpha_{LDA} E_c^{\mathrm{LDA}} + \alpha_{PBEsol} E_c^{PBEsol} + \alpha_{nl} E_c^{\mathrm{nl}},$$

where the none-local (nl) correlation is of the vdW-DF2 type, similar to BEEF-vdW. We now have a free parameter for each part of the correlation functional, and thus have a more free form than the parametrization in the BEEF-vdW study, and we use PBEsol correlation instead of the PBE correlation.[1]

In total we have 67 parameters in the exchange-correlation model.

8.2 The training datasets

For the mBEEF-vdW functional we choose to use only training datasets previously introduced in the mBEEF and BEEF-vdW studies:

RE42 42 reaction energies to represent gas-phase chemistry.

CE27 27 chemisorption energies with references as in the mBEEF study.

Sol54Ec Cohesive energies of 54 solids.

Sol58LC The derivatives of the cohesive energies with respect to the crystal volumes around the experimental equilibrium taken from the Sol58LC dataset.

S22x5 Non-covalent interaction of the 22 intermolecular interaction energies, with the interaction energies of the relative distances of 0.9, 1.0, 1.2, 1.5 and 2.0 as compared to the S22 dataset, which were also used in the BEEF-vdW study.

It should be noted that in this study we chose not optimize directly for the G3/99 dataset. The reason is that we previous observed that the G3/99 fitted very well in BEEF-vdW and mBEEF, but that the description of atomization energies did not carry over to relevant quantities; secondly we find that the relevant information of the set is captured in the reaction energy dataset of RE42.

[1] For LDA see Perdew and Wang [1992], for PBEsol see Perdew et al. [2008] and for vdW-DF2 see Lee et al. [2010].

In this study, the structural geometries and electronic densities were initially that of PBEsol, and then to the first mBEEF-vdW functional in the series of towards a mBEEF-vdW functional fitted to on self-consistent structural geometries and electronic densities.

8.3 Model selection

To find the optimum model for all the datasets we will now use a two step procedure:

First, we fit the individual datasets, where two new concepts will be introduced. We will use the Bayesian error estimation (BEE) ensemble Covariance estimation to rescale the system errors internally in the datasets, and secondly we will be a robust MM-estimator for the loss function to make the optimum model resistant to outliers. We will provide assessment of including each of these two new extensions to the loss function.

Secondly, we will fit to all the datasets simultaneously using a model compromise loss function. In this step all the data points will carry over weights from the individual fits from the BEE ensemble covariance estimation, and from the robust loss function. To find the optimal model complexity, we will introduce the hierarchical cluster bootstrap resampling with the model compromise loss function.

8.3.1 Rescaling datasets from the Bayesian error estimation

It was noted in the machine learning chapter that the Bayesian error estimation ensemble could be used to create a covariance matrix for a dataset, see equation 4.37. We here propose to use the estimated covariance to scale all the data points of a datasets, by transforming the design matrix $\boldsymbol{X}$ with the $\boldsymbol{T}_{BEE}$, where

$$\boldsymbol{T}_{BEE} = \frac{\sqrt{Cov_{BEE}^{-1}}}{\frac{1}{N_d}\mathrm{tr}(\sqrt{Cov_{BEE}^{-1}})}. \tag{8.1}$$

$\boldsymbol{T}_{BEE}$ is found self-consistently, which means that a solution provides $\mathbf{T}_{BEE}$ scaling, that then again will give rise to a slightly changes solution. This procedure is highly convergent and only about a couple of iterations are needed. We will for short refer to the method as BEE scaling or BN, for BEE normalization.

8.3.2 What are outliers in the the dataset

We propose that outliers in the data should be disregarded in the fitting procedure. An outlier is a data point that falls outside of what can be regarded as normal data. This makes us suspicious for that there could be problems associated with the data, and that our model will become less accuracy when we use it on data points outside the training dataset. Here follows some reasons for why outliers can arise in our training datasets:

- Experimental reference value can have a high amount of uncertainty associated with it, and the theoretical model system can be a bad representation of the experimental system.

- The model system calculations can use inadequate setups/potentials/basis sets. The model system can also be non-converged in structure geometries, electronic density or total energy.

- The model space can be incomplete in a way that makes it hard for the model to give descriptions to certain materials properties. Some of the challenges to semi-local functionals are: strong correlation effects, relativistic effects, self-interaction, and long range dispersion. .

For the outliers we cannot know if some of the above effects are in play, so we must rely on robust fitting methods that are resistant to the outliers. The Robust fitting theory offers a theoretical foundation for how to identify outliers and remove them.

The robust MM-estimator procedure

To make our fitting procedure robust, we will use the MM-estimator loss function instead of the least square (LS) loss function that so far have been used[2].

The implementation will be based on Maronna [2011], where the MM-estimator loss function was combined with ridge regression. Following the definition of the MM-estimator in section 4.6.3, we need to choose the constants for the MM-estimator cost function, and we need to have a scheme for which to come up with an initial robust solution. The following sections will provide these constants following Maronna [2011] and how we have chosen to come up with initial robust estimates of the solution.

We will use the method of Maronna [2011] to find the robust solutions for a given regularization. To find the model compromise we however rely on the cross validation methods that have been presented in 4.4 and 4.5.4. The details for how the optimal model complexity is found with the BEE scaling and MM-loss function will follow after the implementation of the MM procedure.

[2] Following the discussion of section 4.6 it was found that the LS loss function is not robust.

8.3.3 Constants for the MM procedure[3]

Recall the cost function for the MM estimator in equation 4.6.3:

$$C(\hat{\boldsymbol{a}}, \omega) = \hat{\sigma}_{init}^2 \sum_{i}^{N_d} \rho\left(\frac{r_j(\hat{\boldsymbol{a}})}{\hat{\sigma}_{init}}\right) + R(\boldsymbol{a}, \omega).$$

To use the MM-estimator cost function we need to define the ρ and the initial robust estimate $\hat{\sigma}_{init}$, that is estimated from ρ_0. For ρ and ρ_0 ρ-functions we use the TukeyBisquare defined in equation 4.66 by $\rho_{bis}(t) = \min\{1-(1-t^2)^3, 1\}$. We furthermore need to choose the c_0 in $\rho_0(t) = \rho_{bis}\left(\frac{t}{c_0}\right)$ for the initial M-estimator, and c for $\rho(t) = \rho_{bis}\left(\frac{t}{c}\right)$, where $c_0 \quad \leq \quad c$.

To the initial scale we want to achieve a high BDP. We solve for the scale using $\delta = 0.5\left(1 - \frac{N_{eff}}{N_d}\right)$, and the constant c_0 is in Maronna [2011] by

$$c_0 = 7.8464 - 34.6565 \cdot \delta + 75.2573 \cdot \delta^2 - 62.5880 \cdot \delta^3, \qquad (8.2)$$

where the relationship between c_0 to δ for illustration is plotted in the margin figure (figure 8.1).

For each initial solution, $\boldsymbol{a}_{init}$, c_0 is thus found by the above equation, and used to calculate the initial robust estimator $\hat{\sigma}_{init}$.

For the MM estimator, c is chosen as $c = 3.44$ to provide a normal efficiency of 85%.[4]

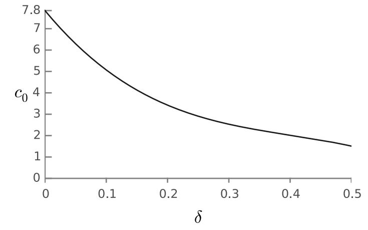


Figure 8.1: The relationship between c_0 and δ.

We furthermore employ a correction to $\hat{\sigma}_{ini}$ and c for high dimensional data, where the number of effective parameters are comparable to the number of data points in the dataset. The correction to $\hat{\sigma}_{ini}$ is given as

$$\tilde{\sigma}_{ini} = \frac{\hat{\sigma}_{ini}}{1-(k_1 + k_2/n)N_{eff}/N_d}, \quad with\, k_1 = 1.29,\, k_2 = -6.02, \qquad (8.3)$$

and we adjusted c to $c \;=\; 4$, when $\hat{N}_{eff}/N_d \;>\; 0.1$.[5]

The IRWLS weights to the dataset will change the balance between the loss term and the regularization term in the cost function, and thus change the effective number of parameters at a given regularization strength. A correction to the regularization strength is therefore introduced, so that RR-MM match the RR-LS solution for a given regularization strength. The correction is given as $\omega' = \sqrt{3\omega/c^2}$.

[3] The approach here follows Maronna [2011].

[4] It was found in Maronna [2011] that a higher efficiency would also lead to a higher unwanted bias of the solution.

[5] See more details in Maronna [2011].

8.3.4 The initial estimator

The initial estimator of the scale is found, following Maronna [2011], using the S-estimator, which is given by the cost function

$$C(\hat{\boldsymbol{a}}, \omega) = \hat{\sigma}^2 \sum_{j=1}^{N} \rho_0 \left(\frac{r_j(\hat{\boldsymbol{a}})}{\hat{\sigma}} \right) + R(\omega),$$

where $\hat{\sigma}$ is now no longer fixed as in the RR-MM cost function of equation. The solution to the RR-SE problem is again found using the IRWLS procedure.
To initialize the algorithm, $N_{bootstrap}$ bootstrap resampled datasets are created, and the solution for each of these are found. The is similar to using the trimmed least squares loss function. For each of these solutions the robust initial scale is fund to the full dataset, but only the N_{keep} lowest estimator solutions is fully iterated to their solution using the IRWLS for the RR-SE estimator.

From the N_{keep} RR-SE solutions we select the one with the lowest robust RR-SE scale $\hat{\sigma}$, and this is now the $\hat{\sigma}_{init}$ scale in the RR-MM procedure, after we have applied the correction of equation 8.3.

The number of effective parameters for the S-estimator is calculated similarly to the RR-MM method. We use $N_{bootstrap} = 500$, and $N_{keep} = 50$, which we have found provide stable results for our datasets.

8.3.5 Defining the loss function

In the following we will propose four different loss functions that we wish to compare:

RR-LS The standard Least Squares loss function

RR-LS-BN The standard Least Squares loss function where the datasets has been scaled self-consistently with the Bayesian covariance-scaling

RR-MM The robust MM-estimator loss function as given above combined

RR-MM-BN The robust MM-estimator loss function as given above combined, where the datasets has been scaled self-consistently with the Bayesian covariance-scaling

These are all combined with the same regularization prior for a full cost function of their respective names above.

The regularization is defined by the smoothness to the exchange enhancement factor as in mBEEF and with additional regularization on the correlation parameters, where the regularization strength to the correlation parameters are scaled by 10^{-3} to the overall regularization strength. In all model optimizations of this chapter, we use have that the origo of the exchange solution is given as $F_x = 1$ and the origo of the correlation solution is given as $\alpha_{LDA} = \alpha_{PBEsol} = 0.5$, and $\alpha_{nl} = 1$.

The optimal model complexity is found with the Bootstrap.632 estimator using one set of samples. For the MM loss function, we update the robust IRWLS weights to the dataset in the Bootstrap.632 regularization minimum, and repeat until convergence in the model complexity and the IRWLS weights. Convergence is usually reached within 3 iterations. In the bootstrap estimated prediction error, the IRWLS and Bayesian error estimation scaling weights are carried over, such that the model complexity optimization is internally consistent.

The fitting procedure for fitting multiple datasets is the following. For the BN or MM loss functions the weights on the individual systems from these two approaches combines is carried over to the compromise fit. For the compromise fit we define the loss function with the geometric mean, similar to the mBEEF study. The optimal model is then found by minimizing the estimated prediction error found with the bootstrap.632, where we use the geometric mean loss function for the datasets and the hierarchical resampling; as describe in the machine learning chapter. For the estimated prediction error for the model compromise, we similar to the for the individual fits carry over the weights from the IRWLS procedure and T_{BEE} weights to the estimating prediction error.

8.4 Outlier detection in RE42

We want to test the procedure on a realistic test dataset, where the result of the procedure can easily be evaluated. We therefore introduce a number of outlier data points in the RE42 dataset, and compare the RR-LS and RR-MM-BN cost function solutions. The outliers are made by adjusting target values in the target vector. We shift the first two targets by -1 eV and the next two by +1 eV. We expect a lower model complexity from the RR-LS when outliers are introduced, as the outliers will lead to less transferable models within the dataset in the bootstrap samples.

In the figure below (figure 8.2) we compare the regularization curves for the RE42 dataset with outliers, using the RR-LS and RR-MM methods. We find that the RR-MM minimizes EPE for a more complex model, and furthermore provides a less shallow minimum than RR-LS.

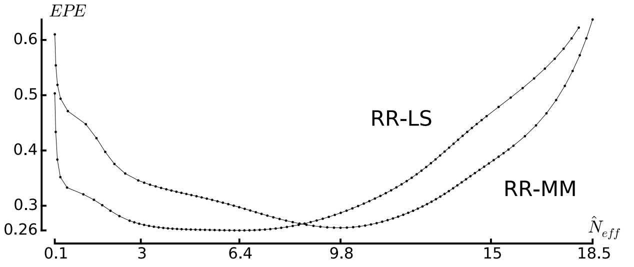


Figure 8.2: *Bootstrap*$_{.0632}$ Estimated prediction error (EPE) versus the number of effective parameters (N_{eff}) for the RE42 dataset with introduced outliers with the RR-LS and RR-MM cost functions. The RR-MM minimum is less shallow, and the model has 9.8-6.4=3.4 more effective parameters then the optimal model for the RR-LS method.

Next we will compare the root mean square deviation (RMSD) for the data points in the RE42 dataset omitting the first 4 data points: No outliers: RR-LS: 13 meV ($N_{eff} = 8.7$), RR-MM: 8.0 meV ($N_{eff} = 9.7$); With outliers: RR-LS: 44 meV ($N_{eff} = 6.4$), RR-MM: 9.7 meV ($N_{eff} = 9.8$).

We find that the RR-MM method provides a more complex model for RE42 in general, and is only slightly affected by the outliers with respect to the model complexity and the RMSD for the good data points. The larger model complexity for RR-MM over RR-LS, when there is no outliers introduced, propose that some of the data in the dataset are not well-defined within the model space.[6]

The study shows that the RR-MM cost function with BEE scaling is indeed more resistant to outliers than the RR-LS cost function.[7]

[6] Note that results of this section will not correspond to later results for the RR-MM method on RE42, as the MM procedure was slightly adjusted later to make the BDP of the method similar to that of Maronna [2011].

[7] The BEE scaling is the same for the outliers with and without the outliers, so the robustness is from the MM-estimator solely.

8.5 Assessment of best loss functions for the exchange correlation model selection

We will now compare the different models that has been proposed by first fitting them to the individual datasets, and in the following making the compromise to all datasets. We follow the methodology for fitting with the MM loss function and with the Bayesian scaling given previous in both instances. For comparing the loss functions, we will introduce the introduce the tilde version of the ERR, err and EPE for the Bootstrap procedure, where the IRWLS and $\boldsymbol{T}_{BEE}$ weights have not been introduced. These estimates are thus comparable between the methods, and $\widetilde{ERR}$ provide a direct measure for the transferability of the method, which we optimize for. For the individual fits $\widetilde{ERR}$ refer to that of the standard Bootstrap.632 method and for the model compromise fit $\widetilde{ERR}$, $\widetilde{err}$ and $\widetilde{EPE}$ are that of the hierarchical bootstrap with geometric mean loss function.

We present two sets of data where the structural geometries and the electronic densities are that of respectively: In run 0 the PBEsol; and in run 1 a mBEEF-vdW functional fit with the RR-MM method as outlined here but with the c_0 fixed to that of a 85% efficiency, which we expect only have a minor influence on the fit.

With data from both runs provided, one can get an understanding of the influence that self-consistency has. However, to fully assess this, one would have to compare the non-self-consistent results with the self-consistent results for the same model. This has not been included as we do not have data for all the runs towards self-consistency, and we would therefore not be able to make the analysis fully.

The best loss function should however be independent of the density that is used, as we are evaluating it on the parametrization energies for the same density as it was fitted to.

8.5.1 Comparing loss functions for the individual datasets

For the model optimization for the individual datasets we compare the following quantities: The effective number of parameters (N_{eff}); The transferability estimate $\widetilde{ERR}$ which that of $\sqrt{\widehat{ERR}}$ of the Bootstrap .632 procedure without the scaling of IRWLS and $\boldsymbol{T}_{BEE}$, following the discussion above; RMSD, MAD and MSD are the root mean square, mean absolute and mean signed deviation for the model for the unscaled data points; min, and max are the minimum and maximum values of the unscaled data points; $\boldsymbol{r}$ are the deviations for the model prediction to the target values for all data points; $\hat{\boldsymbol{r}} = \boldsymbol{T}_{BEE}\boldsymbol{r}$

(the Bayesian error estimation scaled deviations); $\boldsymbol{w}$ shows the IRWLS for all the data points; and for the last three stats it is noted that N_d is the number of data points.

CE27

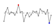
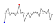
	N_{eff}	$\widetilde{ERR}$	RMSD	MAD	MSD	min	max	$\boldsymbol{r}$	$\bar{\boldsymbol{r}}$	$\boldsymbol{w}$	$\sum w_i$	$\sum w_i/N_d$	$N_{eff}/\sum w_i$
run 0:													
RR-LS	4.8	227	162	122	-12	-412	254				27.0	1.00	0.18
RR-LS-BN	5.4	239	172	127	-16	-438	349				27.0	1.00	0.20
RR-MM	5.1	198	180	126	-55	-491	194				22.0	0.81	0.23
RR-MM-BN	5.7	208	167	117	-19	-424	387				23.9	0.88	0.24
run 1:													
RR-LS	8.7	300	140	111	-0	-242	330				27.0	1.00	0.32
RR-LS-BN	6.1	328	207	156	-12	-601	522				27.0	1.00	0.23
RR-MM	8.7	275	142	111	2	-243	343				26.1	0.97	0.33
RR-MM-BN	6.3	313	218	166	-2	-365	624				24.7	0.92	0.26

Table 8.1: CE27 model statistics with deviations in meV.

For chemisorption dataset (CE27), see figure 8.1: We highlight that the all models become more complex for run 1 compared to run 0. For this dataset we have structure geometry optimization, so we expect that there would be a larger difference between run 0 and run 1. The data is observed as more noisy when the densities are not self-consistent, and the result is similar to when outliers were added to the RE42 dataset. There are no major outliers in run 1.

The RR-MM loss function has the lowest $\widetilde{ERR}$ on both run 0 and run 1.

RE42

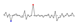
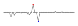
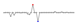
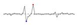
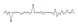
	N_{eff}	$\widetilde{ERR}$	RMSD	MAD	MSD	min	max	$\boldsymbol{r}$	$\bar{\boldsymbol{r}}$	$\boldsymbol{w}$	$\sum w_i$	$\sum w_i/N_d$	$N_{eff}/\sum w_i$
run 0:													
RR-LS	14.0	181	58	41	-1	-178	148				42.0	1.00	0.33
RR-LS-BN	13.4	195	87	54	4	-189	370				42.0	1.00	0.32
RR-MM	13.3	442	446	154	-52	-1983	1757				33.2	0.79	0.40
RR-MM-BN	10.5	290	279	121	14	-881	1142				32.3	0.77	0.32
run 1:													
RR-LS	14.9	198	62	44	-4	-224	142				42.0	1.00	0.35
RR-LS-BN	14.9	189	79	52	7	-179	263				42.0	1.00	0.35
RR-MM	15.9	161	70	35	-11	-380	109				38.6	0.92	0.41
RR-MM-BN	15.2	181	78	44	-4	-310	276				38.9	0.93	0.39

Table 8.2: RE42 model statistics with deviations in meV.

For reaction energies dataset (RE42), see figure 8.2: The different between run 0 and run 1 in the number of effective parameters is large, especially for the RR-MM methods. It can also be observed that the number of outliers are drastically reduced. The biggest outlier that persist in run 1 is that of the reaction $O_2 + 2H_2 \rightarrow 2H_2O$.

The RR-MM again outperforms the other loss functions, but with an overall statistic that is worse than RR-LS (RMSD 62 to 70 meV).

Sol54Ec

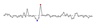
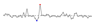
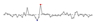
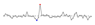
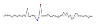
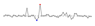
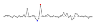
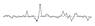
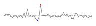
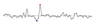
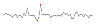
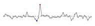
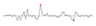
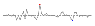
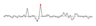
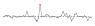
	N_{eff}	$\widetilde{ERR}$	RMSD	MAD	MSD	min	max	$\boldsymbol{r}$	$\bar{\boldsymbol{r}}$	$\boldsymbol{w}$	$\sum w_i$	$\sum w_i/N_d$	$N_{eff}/\sum w_i$
run 0:													
RR-LS	9.8	313	233	142	1	-506	1135				54.0	1.00	0.18
RR-LS-BN	10.4	318	228	146	-14	-552	1011				54.0	1.00	0.19
RR-MM	12.1	255	240	130	36	-459	1217				42.8	0.79	0.28
RR-MM-BN	11.7	255	238	131	29	-460	1201				41.7	0.77	0.28
run 1:													
RR-LS	9.8	316	233	142	1	-506	1135				54.0	1.00	0.18
RR-LS-BN	10.6	320	227	145	-14	-560	995				54.0	1.00	0.20
RR-MM	11.0	282	278	158	23	-517	1250				36.7	0.68	0.30
RR-MM-BN	11.9	254	238	131	29	-463	1194				41.9	0.78	0.28

Table 8.3: Sol54Ec model statistics with deviations in meV.

For cohesive energies solids dataset (Sol54Ec), see figure 8.3: The number of effective parameters decreases for the RR-MM loss functions for run 1 compared to run 0. There are a large number of outliers in the dataset. For this dataset the RR-LS method is the best performing in the $\widetilde{ERR}$ estimate. Looking at the $\boldsymbol{r}$ plots we see a few very big outliers in the dataset, and a big chuck that we are able to describe exceedingly well. The data of Sol54Ec is experimental, and the fits suggest that one should take a closer look at the underlying data.

Sol58LC

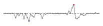
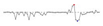
	N_{eff}	$\widetilde{ERR}$	RMSD	MAD	MSD	min	max	$\boldsymbol{r}$	$\bar{\boldsymbol{r}}$	$\boldsymbol{w}$	$\sum w_i$	$\sum w_i/N_d$	$N_{eff}/\sum w_i$
run 0:													
RR-LS	11.8	24	15	13	2	-41	43				58.0	1.00	0.20
RR-LS-BN	7.0	24	20	14	-1	-49	65				58.0	1.00	0.12
RR-MM	8.4	28	28	15	-7	-91	100				32.0	0.55	0.26
RR-MM-BN	11.8	28	27	15	-6	-89	88				43.9	0.76	0.27
run 1:													
RR-LS	12.0	24	15	13	2	-41	42				58.0	1.00	0.21
RR-LS-BN	7.0	24	20	14	-1	-49	65				58.0	1.00	0.12
RR-MM	12.1	29	29	16	-8	-86	87				38.9	0.67	0.31
RR-MM-BN	8.8	29	29	16	-8	-87	100				31.1	0.54	0.28

Table 8.4: Sol58LC model statistics with deviations of the cohesive energies at the experimental lattice parameter (meV/cubic angstrom) .

For the derivative to the cohesive in the experimental lattice parameter (Sol58LC), see figure 8.4: The outliers in the dataset for the two runs and for both the MM loss functions looks similar to that of the cohesive energies. The effective number of parameters change change for RR-MM and RR-MM-BN method in opposite directions between run 0 and run 1; for RR-MM N_{eff} goes up and for RR-MM-BN N_{eff} goes down. The LS loss functions outcompete the MM in the $\widetilde{ERR}$ estimate.

For the S22x5 dataset in the 1.0 distance, see figure 8.5: For this and for the other of S22x5 the MM method gives very similar results and the difference is mostly in the BN scaling, which has a slightly lower $\widetilde{ERR}$ estimate. $N_{eff}/\sum \boldsymbol{w}_i$ figure is fairly high, which will s22x5-1.0

	N_{eff}	$\widetilde{ERR}$	RMSD	MAD	MSD	min	max	$\boldsymbol{r}$	$\bar{\boldsymbol{r}}$	$\boldsymbol{w}$	$\sum w_i$	$\sum w_i/N_d$	$N_{eff}/\sum w_i$
run 0:													
RR-LS	9.2	29	9	7	-1	-21	18				22.0	1.00	0.42
RR-LS-BN	11.2	28	7	5	-1	-17	13				22.0	1.00	0.51
RR-MM	9.4	28	9	6	-1	-23	19				21.6	0.98	0.43
RR-MM-BN	11.7	31	6	5	-0	-14	12				21.9	1.00	0.53
run 1:													
RR-LS	10.9	27	6	5	-0	-13	11				22.0	1.00	0.49
RR-LS-BN	8.5	28	14	9	-2	-47	15				22.0	1.00	0.39
RR-MM	10.6	27	7	5	-0	-13	11				21.7	0.99	0.49
RR-MM-BN	8.5	28	14	9	-2	-46	15				21.8	0.99	0.39

Table 8.5: S22x5 1.0 model statistics with deviations in meV .

make the c_0 value go up, see figure 8.1. The dataset is therefor so small that a data point needs to be more extreme for the MM method to judge it as an outlier. For the 0.9 distance, see figure 8.6, one outlier is however observed, and the MM methods outperform the LS methods in $\widetilde{ERR}$. For completion the loss function comparison statistics 1.2, 1.5, and 2.0 parts of the S22x5 dataset are provided in figure 8.7, 8.8 and 8.9, and we note that the difference between the MM and LS loss function is small because no outliers could be identified.

s22x5-0.9

	N_{eff}	$\widetilde{ERR}$	RMSD	MAD	MSD	min	max	$\boldsymbol{r}$	$\bar{\boldsymbol{r}}$	$\boldsymbol{w}$	$\sum w_i$	$\sum w_i/N_d$	$N_{eff}/\sum w_i$
run 0:													
RR-LS	8.6	31	12	10	-1	-30	18				22.0	1.00	0.39
RR-LS-BN	9.0	29	12	8	-1	-38	21				22.0	1.00	0.41
RR-MM	8.4	29	13	10	-2	-40	19				21.4	0.97	0.39
RR-MM-BN	9.3	29	12	8	0	-38	16				21.7	0.99	0.43
run 1:													
RR-LS	8.0	29	13	10	-1	-34	19				22.0	1.00	0.36
RR-LS-BN	8.5	28	14	10	-2	-47	16				22.0	1.00	0.38
RR-MM	10.3	26	17	7	-4	-77	13				20.7	0.94	0.50
RR-MM-BN	8.9	26	17	9	-4	-67	13				21.2	0.96	0.42

Table 8.6: S22x5 0.9 model statistics with deviations in meV .

s22x5-1.2

	N_{eff}	$\widetilde{ERR}$	RMSD	MAD	MSD	min	max	$\boldsymbol{r}$	$\bar{\boldsymbol{r}}$	$\boldsymbol{w}$	$\sum w_i$	$\sum w_i/N_d$	$N_{eff}/\sum w_i$
run 0:													
RR-LS	10.4	23	5	4	-1	-9	14				22.0	1.00	0.47
RR-LS-BN	9.5	21	10	5	-1	-33	27				22.0	1.00	0.43
RR-MM	11.2	23	5	4	-0	-7	13				21.9	1.00	0.51
RR-MM-BN	10.0	22	9	5	-1	-31	27				21.8	0.99	0.46
run 1:													
RR-LS	9.0	17	6	5	-1	-11	12				22.0	1.00	0.41
RR-LS-BN	8.2	17	9	6	-0	-28	15				22.0	1.00	0.37
RR-MM	8.7	17	6	5	-1	-12	13				21.5	0.98	0.40
RR-MM-BN	8.4	18	9	6	-0	-26	15				21.5	0.98	0.39

Table 8.7: S22x5 1.2 model statistics with deviations in meV .

s22x5-1.5

	N_{eff}	$\widetilde{ERR}$	RMSD	MAD	MSD	min	max	$\boldsymbol{r}$	$\bar{\boldsymbol{r}}$	$\boldsymbol{w}$	$\sum w_i$	$\sum w_i/N_d$	$N_{eff}/\sum w_i$
run 0:													
RR-LS	10.8	9	2	2	-0	-5	3				22.0	1.00	0.49
RR-LS-BN	10.1	11	3	2	-0	-8	9				22.0	1.00	0.46
RR-MM	10.8	9	2	2	-0	-5	3				21.8	0.99	0.49
RR-MM-BN	10.3	10	3	2	-0	-6	8				21.7	0.99	0.47
run 1:													
RR-LS	7.7	7	3	2	-0	-4	6				22.0	1.00	0.35
RR-LS-BN	7.0	9	5	4	-0	-16	7				22.0	1.00	0.32
RR-MM	8.2	8	3	2	-0	-4	6				21.6	0.98	0.38
RR-MM-BN	6.9	9	5	3	-0	-15	6				21.6	0.98	0.32

Table 8.8: S22x5 1.5 model statistics with deviations in meV .

s22x5-2.0

	N_{eff}	$\widetilde{ERR}$	RMSD	MAD	MSD	min	max	$\boldsymbol{r}$	$\bar{\boldsymbol{r}}$	$\boldsymbol{w}$	$\sum w_i$	$\sum w_i/N_d$	$N_{eff}/\sum w_i$
run 0:													
RR-LS	7.7	6	2	2	-1	-3	2				22.0	1.00	0.35
RR-LS-BN	8.0	8	4	3	1	-10	14				22.0	1.00	0.36
RR-MM	8.5	6	2	1	-1	-3	2				21.7	0.98	0.39
RR-MM-BN	7.9	8	4	3	1	-10	12				21.4	0.97	0.37
run 1:													
RR-LS	11.2	6	1	1	0	-2	3				22.0	1.00	0.51
RR-LS-BN	10.8	5	2	1	0	-5	3				22.0	1.00	0.49
RR-MM	11.9	5	1	1	0	-2	3				21.9	1.00	0.54
RR-MM-BN	10.8	5	2	1	0	-5	3				21.9	1.00	0.49

Table 8.9: S22x5 2.0 model statistics with deviations in meV .

Overall for the individual datasets, we saw that the MM loss function outperformed the LS method for CE27 and RE42 in the $\widetilde{ERR}$ estimate, and only a small number of outliers were observed for these datasets in run 1. For the solids datasets the LS method performed better than the MM methods, and these datasets had many outliers each, both in run 0 and run 1. Very few outliers were observed in the s22x5 dataset, and the difference between the MM and LS method.

The BN transformation resulted in higher $\widetilde{ERR}$ for the CE27 and RE42 datasets and a lower $\widetilde{ERR}$ for the Sol54Ec dataset. The transformation did not affect the results for the s22x5 datasets much for the $\widetilde{ERR}$ estimate.

The model compromise solution

For the compromise loss function we choose the following $\breve{\boldsymbol{w}}$: For the s22x5 datasets (0.9, 1.0, 1.2, 1.5, 2.0 bonding length scales) $\breve{w} = 0.2$, such that the combined weight of the s22x5 datasets is 1; For CE27 and RE42 we set $\breve{w} = 2$, to give importance to these datasets; Lastly, for Sol54Ec and Sol58LC $\breve{w} = 1$. With these weights, we seek a generally applicable functional especially suited for heterogeneous catalysis studies.

Table 8.10: Comparing the loss functions for the model compromise.

	Model compromise			
	N_{eff}	$\widetilde{ERR}$	$\widetilde{err}$	$\widetilde{EPE}$
run 0:				
RR-LS	11.2	365	121	547
RR-LS-BN	13.1	435	122	595
RR-MM	10.9	291	135	495
RR-MM-BN	6.7	343	157	538
run 1:				
RR-LS	10.0	352	141	541
RR-LS-BN	12.0	367	140	551
RR-MM	10.5	295	140	499
RR-MM-BN	7.8	322	197	531

The overall comparison statistics can be seen in figure 8.10, with the $\widetilde{ERR}$, $\widetilde{err}$ and $\widetilde{EPE}$, where the weights $\breve{\boldsymbol{w}}$ have been taken into account in the geometric mean estimates of the datasets. We find that the RR-MM loss function performs best for both run 0 and run 1.

The exchange enhancement factor for the RR-MM fit of run 1 is shown figure 8.3. The exchange enhancement factor breaks the LDA limit and is first very flat for then to follow the MS0 functional approximatively along both the s and α dimensions plotted here.

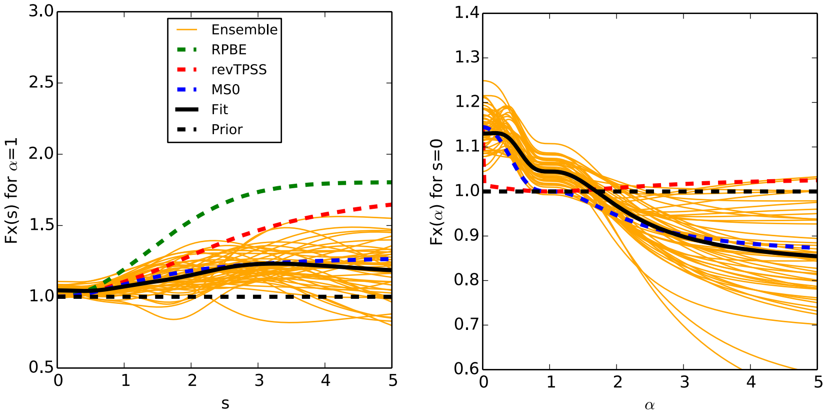


Figure 8.3: The optimal model for the mBEEF-vdW run 1 densities (black) with a Bayesian ensemble of exchange models (yellow).

The correlation parameters for the functional are given as $\alpha_{LDA} = 0.43 \pm 0.16$, $\alpha_{PBEsol} = 0.39 \pm 0.16$, and $\alpha_{nl} = 0.87 \pm 0.15$, where the second number is one standard deviation of the Bayesian error estimation ensemble for the parameter. It notably follows the MS0 functional in the exchange enhancement factor, but it breaks the LDA limit for $s = 0, \alpha = 1$.

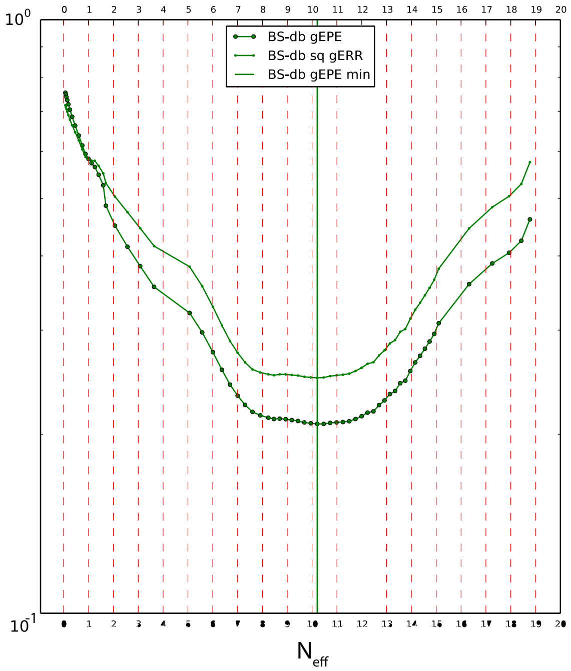


Figure 8.4: A representative regularization curve for the RR-MM method on run 1.

In figure 8.4 we have plotted a representative regularization plot for MM-LS run 1.[8] The minimum in the EPE curve is very well defined, in contrast to the cluster-LOOCV used in the mBEEF study.

[8] It is not the exact plot from above as the robust weighting carries a bit of variation between runs, and the robust weight have are not the exact of the solution given above.

8.5.2 Benchmark

For the RR-MM solution we provide non-self-consistent benchmark in the table 8.11 and table 8.12, where the functional of RR-MM run 1 is referred to as mBEEF-vdW RC2, or for short mBEEF-vdW. The functionals have been divided in groups as of their model complexity, hence from the top: GGA; MGGA; GGA+vdW; and with the mBEEF-vdW MGGA+vdW. These functionals have been selected as the most representative functionals for each group that are accessible in our code.[9]

We do not have reference data for the derivative to the Sol58LC dataset at the experimental lattice parameters, so it is here omitted. We expect that the non-self-consistent results resembles the self-consistent results to a very high degree, as the functional is close to the functional that was used to create the electronic densities and structural geometries.

In the following account for the benchmark of mBEEF-vdW, we evaluate the functional as a general applicable surface science functional that is capable of capturing non-covalent bonding.

For the S22x5 benchmark, in table 8.11, we use the geometric mean of the different lengths to rank the functionals. The mBEEF-vdW is by the geometric mean of RMSD and MAD, standards performing in between the optPBE-vdW and optB88-vdW; both of which have been optimized for the S22 set, and a little worse than the C09-vdW functional.

	s22x5 0.9			s22x5 1.0			s22x5 1.2			s22x5 1.5			s22x5 2.0			s22x5 GM		
(meV)	MAD	MSD	RMSD	MAD	MSD	RMSD	MAD	MSD	RMSD	MAD	MSD	RMSD	MAD	MSD	RMSD	MAD	MSD	RMSD
PBE	160	160	230	120	120	163	64	64	80	27	27	32	9	9	12	50	50	65
PBEsol	100	51	134	79	59	113	51	47	69	25	25	31	9	9	12	39	32	52
RPBE	320	320	405	224	224	275	102	102	122	33	33	42	9	9	12	74	74	92
TPSS	205	205	274	160	160	207	87	87	105	35	35	42	10	10	13	64	64	80
revTPSS	184	184	232	146	146	182	84	84	100	35	35	43	11	11	15	62	62	77
oTPSS	273	273	353	208	208	260	106	106	125	39	39	48	11	11	13	75	75	94
MS0	109	109	150	79	79	105	45	45	57	23	23	28	9	9	12	38	38	50
mBEEF	100	96	141	60	54	80	23	16	29	14	9	18	7	7	9	27	22	36
vdW-DF	140	140	168	71	70	89	32	4	40	15	-13	20	4	-4	5	28	17	36
vdW-DF2	99	99	123	44	43	55	13	5	16	4	2	5	5	5	6	16	12	20
optPBE-vdW	31	29	40	20	-1	24	28	-25	38	20	-20	24	5	-5	7	17	9	22
optB88-vdW	19	17	22	13	5	15	13	-4	18	6	-3	9	2	1	3	8	4	11
C09-vdW	21	-13	30	13	-3	17	13	-3	16	11	-6	14	2	-2	3	10	4	13
BEEF-vdW	136	136	170	74	72	93	27	6	35	6	-5	8	3	2	4	23	14	28
mBEEF-vdW RC2	30	11	42	29	0	37	17	-5	21	6	-2	7	4	3	6	13	2	17

Table 8.11: Benchmark for the s22x5 sets. The geometric mean (GM) to the S22x5 statistics has been added in the right. The mBEEF-vdW RC2 are non-self-consistent, while all other references have been calculated self-consistently.

[9] All references excluding the mBEEF and the mBEEF-vdW are given here:
PBE: Perdew et al. [1996a], PBEsol: Perdew et al. [2008]; RPBE:Hammer et al. [1999]; TPSS: Tao et al. [2003]; revTPSS: Perdew et al. [2009]; oTPSS: Goerigk and Grimme [2011]; MS0:Sun et al. [2012b]; vdW-DF Dion et al. [2004]; vdW-DF2 Lee et al. [2010]; optPBE-vdW, optB88-vdW: Klimes et al. [2010]; C09-vdW Cooper [2010]; BEEF-vdW Wellendorff et al. [2012].

	CE27			Sol54Ec			RE42		
(meV)	MAD	MSD	RMSD	MAD	MSD	RMSD	MAD	MSD	RMSD
PBE	701	-682	824	240	-80	315	298	-85	420
PBEsol	1461	-1461	1610	395	377	549	480	-287	728
RPBE	164	32	203	528	-511	616	250	113	333
TPSS	371	-337	469	237	-41	308	250	82	326
revTPSS	413	-395	515	274	105	384	395	203	519
oTPSS	246	-201	340	302	-162	368	246	24	308
MS0	249	43	312	276	-79	380	457	-71	625
mBEEF	180	-56	250	268	-172	400	261	-18	318
vdW-DF	208	-94	255	546	-467	659	394	237	522
vdW-DF2	289	-152	374	544	-542	652	403	236	537
optPBE-vdW	690	-690	795	210	-90	307	268	62	346
optB88-vdW	906	-906	1021	205	24	292	259	17	344
C09-vdW	1324	-1324	1440	352	344	488	327	-111	452
BEEF-vdW	174	-32	201	380	-345	496	290	142	372
mBEEF-vdW RC2	191	-22	243	259	66	400	305	-12	376

Table 8.12: Benchmark. The mBEEF-vdW RC2 are non-self-consistent, while all other references have been calculated self-consistently.

Next we will take a look at the CE27, Sol58Ec and RE42 datasets in table 8.12. For the chemisorption systems of CE27, mBEEF-vdW has a RMSD that is a little higher than that of BEEF-vdW and RPBE, and on the same level as vdW-DF and mBEEF. These functional are however much worse performing to the S22x5. The optB88, optPBE and C09-vdW functionals have RMSD that are 3-4 times that of the mBEEF-vdW. The mBEEF-vdW can therefore like no other functional bridge between the chemisorption and the non-covalent binding of the S22x5 dataset.

For the Sol54Ec dataset the RMSD of mBEEF-vdW is in the range of many of the comparable functional. It is slightly higher than for the optPBE-vdW and optB88-vdW, but better than the vdW-DF, vdW-DF2 and BEEF-vdW. The mBEEF-vdW is in the upper end of the spectrum of the RMSD for the MGGA functional. To the Sol54Ec dataset, we however did not expect to have a high performance within the dataset, as much many of the data points were deemed outliers. We however do not have a objective way of take this into account in the statistics.

For the reaction energies of the RE42 dataset, mBEEF-vdW has RMSD very similar to the optB88-vdW and optPBE-vdW, and lower than the C09 functional. In general the mBEEF-vdW is within reach of functionals with the lowest RMSD on RE42. As for the Sol54Ec dataset, it was observed that the RE42 had a couple of big outliers and these will influence the statistics somewhat.

Overall the benchmark indicates that the mBEEF-vdW is a general applicable surface science functional that is capable of capturing non-covalent bonding. The functional can bridge between the non-covalent binding of S22x5 and the chemisorption like no other functional benchmarked here.

8.6 Summary

In the mBEEF-vdW study presented here, we foremost introduced a number of new machine learning tools to handle some of the issues that had been observed in the BEEF-vdW and mBEEF studies.

For the loss function we introduced the robust fitting scheme to make the optimization procedure resistant to outliers in data, while still be highly efficient for data without outliers. We proved how this could be very beneficial, by introducing artificial outliers in one of our datasets, and the MM estimator loss function proved to handle this much better than the RR loss function previously used.

We introduced an internal scaling procedure BEE scaling, based on the Bayesian error estimation covariance matrix, to scale the datasets to be more internal consistent.

And, we introduced a hierarchical bootstrap resampling cross validation method, and made it compatible with the model compromise loss function for several datasets. This method also showed to perform in the intended manner, by making the regularization curve much more deep, such that overfits can be detected easier.

We tested the methods using the bootstrap transferability error estimator to judge the methods and found that the RR-MM improved transferability notably, but the Bayesian error estimation scaling did not.

The mBEEF-vdW functional was then fitted with the now tested procedures, and a non-self-consistent benchmark indicated that the mBEEF-vdW functional is a general applicable surface science functional that is capable of capturing non-covalent bonding. Full self-consistent calculations is however needed before the assessment can be completed.

9 Future development

We still have many ideas for improvements left to explore within the parts of the BEEF functional development. These improvements could result in more accurate error estimating functional; both at the current computational level and beyond. Some of these ideas for improvements will be provided here.

9.1 Model space

In all the studies so far, only the exchange energy has been fully parametrized. The exchange is the largest contributor to binding of matter in most cases, but the correlation still play a very important role. It would therefore be interesting to explore a full parametrization of the semi-local correlation. Such a parametrization should use the MGGA ingredient, which has currently not been used in the correlation functionals of the BEEF functional, but which is used in correlation functionals for other MGGA functionals.

Another step that could be taken without extra computational burden is to include the VV10 non-local correlation instead of the vdW-DF type non-local correlation. This functional has been shown to yield better statistics than comparable functionals.

Currently, no steps have been taken to remove self-interaction for the functionals in BEEF. There are several procedures for doing; being the Hubbard U method or a form of the PZ-SIC. It is however, a requirement for many applications in heterogeneous catalysis that the methods can be used fully self-consistent, and include force corrections.

Another approach forward would be to include the exact exchange, which would however also make the functional more computational expensive. Either in a partial form or in a screened form. This could also as mentioned for the screened exchange methods, provide a means to remove the self-interaction error.

9.2 Training Datasets

The inclusion of datasets that covers a variety of material properties that exists has been a very important part of developing the three functionals BEEF-vdW, mBEEF and mBEEF-vdW. And to provoke future advancements in the model development, as well as ensure that the coefficients to a more parametrized model can be properly determined, more high quality data is sought. Fortunately is has become easier to get hand on these data, and the inclusion and testing of new varieties of materials properties data should therefore continue.

9.3 Method development

Within the fitting approach presented here there are still many machine learning tools that we have not yet tested. For the exchange parameter space the smoothness regularization seem to have a good physical grounding, and it has been possible to fit a highly parametrized model, and still achieve a high level of transferability, as was shown in all the three studies for the BEEF functionals. However, for a fully parametrized correlation other types of prior estimations could be useful. One way to do this would be to move away from the squared model length regularization cost of ridge regression, similarly to what we have done for the loss function, e.g. would be the lasso method.

A number of non-linear machine learning tools could also provide valuable insight into the functional form of the exchange correlation functional; such as support vector machines, kernel methods or neural networks.

10 Bibliography

C. Adamo and V. Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.*, 110:6158, 1999. 26

A. B. Alchagirov, J. P. Perdew, J. C. Boettger, R. C. Albers, and C. Fiolhais. Energy and pressure versus volume: Equations of state motivated by the stabilized jellium model. *Phys. Rev. B*, 63:224115, 2001. 34, 35

R. Armiento and S. Kümmel. Orbital Localization, Charge Transfer, and Band Gaps in Semilocal Density-Functional Theory. *Physical Review Letters*, 111(3):036402, July 2013. ISSN 0031-9007. DOI: 10.1103/PhysRevLett.111.036402. URL `http://link.aps.org/doi/10.1103/PhysRevLett.111.036402`. 27

S. R. Bahn and K. W. Jacobsen. An object-oriented scripting interface to a legacy electronic structure code. *Comput. Sci. Eng.*, 4:56, 2002. 29

A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098, 1988. 21

A. D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.*, 98:5648, 1993. 25

A. D. Becke. Density-functional thermochemistry .5. Systematic optimization of exchange-correlation functionals. *J. Chem. Phys.*, 107 (20):8554, 1997. 25

A. D. Becke and E. R. Johnson. A density-functional model of the dispersion interaction. *J. Chem. Phys.*, 123(15), 2005. 23

A. D. Becke and E. R. Johnson. Exchange-hole dipole moment and the dispersion interaction: High-order dispersion coefficients. *J. Chem. Phys.*, 124, 2006. 23

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 1 edition, 2006. 39

P. E. Blöchl. Projector augmented-wave method. *Phys. Rev. B*, 50: 17953, 1994. 29

K. S. Brown and J. P. Sethna. Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E*, 68: 021904, 2003. 42, 43

Vivaldo Leiria Campo and Matteo Cococcioni. Extended DFT + U + V method with on-site and inter-site electronic interactions. *Journal of physics. Condensed matter : an Institute of Physics journal*, 22 (5):055602, February 2010. ISSN 1361-648X. DOI: 10.1088/0953-8984/22/5/055602. URL http://www.ncbi.nlm.nih.gov/pubmed/21386347. 27

Matteo Cococcioni and Stefano de Gironcoli. Linear response approach to the calculation of the effective interaction parameters in the LDA+U method. *Physical Review B*, 71(3):035105, January 2005. ISSN 1098-0121. DOI: 10.1103/PhysRevB.71.035105. URL http://link.aps.org/doi/10.1103/PhysRevB.71.035105. 27

V. R. Cooper. Van der Waals density functional: An appropriate exchange functional. *Phys. Rev. B*, 81:161104(R), 2010. 24, 33, 104

G. I. Csonka, J. P. Perdew, A. Ruzsinszky, P. H. T. Philipsen, S. Lebegue, J. Paier, O. A. Vydrov, and J. G. Angyan. Assessing the performance of recent density functionals for bulk solids. *Phys. Rev. B*, 79: 155107, 2009. 21, 34

L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople. Assessment of Gaussian-2 and density functional methods for the computation of enthalpies of formation. *J. Chem. Phys.*, 106:1063, 1997. 31, 32

I. Dabo, A. Ferretti, N. Poilvert, Y. Li, N. Marzari, and M. Cococcioni. Koopmans' condition for density-functional theory. *Phys. Rev. B*, 82(11):115121, 2010. 27

M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist. Van der Waals density functional for general geometries. *Phys. Rev. Lett.*, 92:246401, 2004. 24, 76, 104

J. F. Dobson and T. Gould. Calculation of dispersion energies. *J. Phys.: Condens. Matter*, 24(7):073201, 2012. 23

B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statistical Assoc.*, 78:316, 1983. 48

Pavel D Elkind and Viktor N Staroverov. Energy expressions for Kohn-Sham potentials and their relation to the Slater-Janak theorem. *The Journal of chemical physics*, 136(12):124115, March 2012. ISSN 1089-7690. DOI: 10.1063/1.3695372. URL http://www.ncbi.nlm.nih.gov/pubmed/22462843. 27

J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. A. Hansen, H. H. Kristoffersen, M. Kuisma, A. H. Larsen, L. Lehtovaara, M. Ljungberg, O. Lopez-Acevedo, P. G. Moses, J. Ojanen, T. Olsen, V. Petzold, N. A. Romero, J. Stausholm-Møller, M. Strange, G. A. Tritsaris, M. Vanin, M. Walter, B. Hammer, H. Häkkinen, G. K. H. Madsen, R. M. Nieminen, J. K. Nørskov, M. Puska, T. T. Rantala, J. Schiøtz, K. S. Thygesen, and K. W. Jacobsen. Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method. *J. Phys.: Condens. Matter*, 22:253202, 2010. 29

SL Frederiksen, KW Jacobsen, KS Brown, and JP Sethna. Bayesian ensemble approach to error estimation of interatomic potentials. *Physical review letters*, (October):1–4, 2004. DOI: 10.1103/PhysRevLett.93.165501. URL http://prl.aps.org/abstract/PRL/v93/i16/e165501. 59

L. Goerigk and S. Grimme. A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions. *Phys. Chem. Chem. Phys.*, 13:6670, 2011. 33, 104

L. Grafova, M. Pitonak, J. Rezac, and P. Hobza. Comparative study of selected wave function and density functional methods for noncovalent interaction energy calculations using the extended S22 data set. *J. Chem. Theory Comput.*, 6:2365, 2010. 33, 34

S. Grimme, J. Antony, S. Ehrlich, and H. Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.*, 132: 154104, 2010. 23, 33

A. Gulans, M. J. Puska, and R. M. Nieminen. Linear-scaling self-consistent implementation of the van der Waals density functional. *Phys. Rev. B*, 79:201105(R), 2009. 33

P. Haas, F. Tran, and P. Blaha. Calculation of the lattice constant of solids with semilocal functionals. *Phys. Rev. B*, 79:085104, 2009. 35

B. Hammer, L. B. Hansen, and J. K. Nørskov. Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals. *Phys. Rev. B*, 59:7413, 1999. 21, 64, 76, 104

F. Hanke. Sensitivity analysis and uncertainty calculation for dispersion corrected density functional theory. *J. Comput. Chem.*, 32:1424, 2011. 33

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009. 39, 42, 43, 47, 57

J. Heyd, G. E. Scuseria, and M. Ernzerhof. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.*, 118(18):8207, 2003. 26

PJ Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 1964. URL `http://projecteuclid.org/euclid.aoms/1177703732`. 54

G. H. Jóhannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen, and J. K. Nørskov. Combined electronic structure and evolutionary search approach to materials design. *Phys. Rev. Lett.*, 88 (25):255506, 2002.

P. Jurecka, J. Sponer, J. Cerny, and P. Hobza. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.*, 8:1985, 2006. 33

Kristen Kaasbjerg. Statistical Optimization of Quantum Mechanical Calculations. *Master Thesis, Technical University of Denmark*, June 2005. 60, 61

F. O. Kannemann and A. D. Becke. van der Waals interactions in density-functional theory: Intermolecular complexes. *J. Chem. Theory Comput.*, 6:1081, 2010. 33

C. Kittel. *Indtroduction to Solid State Physics*. John Wiley & Sons, Inc., 8th edition, 2005. 15

J. Klimes, D. R. Bowler, and A. Michaelides. Chemical accuracy for the van der Waals density functional. *J. Phys.: Condens. Matter*, 22: 022201, 2010. 24, 33, 76, 104

J. Klimes, D. R. Bowler, and A. Michaelides. Van der Waals density functionals applied to solids. *Phys. Rev. B*, 83:195131, 2011. 21, 24

W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133, 1965. 18, 19

G. Kresse and D. Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B*, 59:1758, 1999. 29

S. Kristyán and P. Pulay. Can (semi)local density functional theory account for the London dispersion forces? *Chem. Phys. Lett.*, 229(3): 175, 1994. 22

Aliaksandr V Krukau, Gustavo E Scuseria, John P Perdew, and Andreas Savin. Hybrid functionals with local range separation. *The Journal of chemical physics*, 129(12):124103, September 2008. ISSN 1089-7690. DOI: 10.1063/1.2978377. URL http://www.ncbi.nlm.nih.gov/pubmed/19045002. 27

Heather J. Kulik and Nicola Marzari. Accurate potential energy surfaces with a DFT+U(R) approach. *The Journal of Chemical Physics*, 135(19):194105, 2011. ISSN 00219606. DOI: 10.1063/1.3660353. URL http://link.aip.org/link/JCPSA6/v135/i19/p194105/s1&Agg=doi. 27

D. C. Langreth and B. I. Lundqvist. Comment on "Nonlocal Van der Waals density functional made simple". *Phys. Rev. Lett.*, 104(9): 099303, 2010. 25

D. C. Langreth, B. I. Lundqvist, S. D. Chakarova-Kack, V. R. Cooper, M. Dion, P. Hyldgaard, A. Kelkkanen, J. Kleis, L. Kong, S. Li, P. G. Moses, E. Murray, A. Puzder, H. Rydberg, E. Schröder, and T. Thonhauser. A density functional for sparse matter. *J. Phys.: Condens. Matter*, 21(8):084203, 2009. 24

C. Lee, W. Yang, and R. G. Parr. Development of the Colic-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785, 1988. 21

K. Lee, E. D. Murray, L. Kong, B. I. Lundqvist, and D. C. Langreth. Higher-accuracy van der Waals density functional. *Phys. Rev. B*, 82: 081101, 2010. 24, 33, 69, 76, 90, 104

E. H. Lieb and S. Oxford. Improved Lower Bound on the Indirect Coulomb Energy. *Int. J. Quantum Chem.*, 19:427, 1981. 21

R.A. Maronna, D.R. Martin, and V.J. Yohai. *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. Wiley, 2006. ISBN 9780470010921. URL http://books.google.com/books?id=iFVjQgAACAAJ. 54, 55

Ricardo a. Maronna. Robust Ridge Regression for High-Dimensional Data. *Technometrics*, 53(1):44–53, February 2011. ISSN 0040-1706. DOI: 10.1198/TECH.2010.09114. URL `http://www.tandfonline.com/doi/abs/10.1198/TECH.2010.09114`. 55, 56, 92, 93, 94, 96

R. M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, 2004. 15, 28, 29

J. Mortensen, K. Kaasbjerg, S. Frederiksen, J. Nø rskov, J. Sethna, and K. Jacobsen. Bayesian Error Estimation in Density-Functional Theory. *Physical Review Letters*, 95(21):216401, November 2005. ISSN 0031-9007. DOI: 10.1103/PhysRevLett.95.216401. URL `http://link.aps.org/doi/10.1103/PhysRevLett.95.216401`. 60, 61

R. Neumann, R. H. Nobes, and N. C. Handy. Exchange functionals and potentials. *Mol. Phys.*, 87(1):1, 1996. 22

Thomas Olsen and Kristian S Thygesen. Beyond the random phase approximation: Improved description of short range correlation by a renormalized adiabatic local density approximation. 2013. 26

Thomas Olsen, Jun Yan, Jens Mortensen, and Kristian Thygesen. Dispersive and Covalent Interactions between Graphene and Metal Surfaces from the Random Phase Approximation. *Physical Review Letters*, 107(15):1–5, October 2011. ISSN 0031-9007. DOI: 10.1103/PhysRevLett.107.156401. URL `http://link.aps.org/doi/10.1103/PhysRevLett.107.156401`. 26

MR Pederson and JP Perdew. 5 SCIENTIFIC HIGHLIGHT OF THE MONTH. *psi-k.org*, (September 2011):77–100, 2011. URL `http://www.psi-k.org/newsletters/News_109/Highlight_109.pdf`. 27

J. P. Perdew and S. Kurth. Density functionals for non-relativistic Coulomb systems in the new century. In M. Marques, C. Fiolhais, and M. Marques, editors, *A Primer in Density Functional Theory*, page 1. Springer, 2003. 16

J. P. Perdew and K. Schmidt. Jacob's ladder of density functional approximations for the exchange-correlation energy. In V. Van Doren, C. Van Alsenoy, and P. Geerlings, editors, *Density Functional Theory and its Application to Materials*, volume 577 of *AIP Conference Proceedings*, page 1, 2001. 16, 18, 20

J. P. Perdew and Y. Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B*, 45:13244, 1992. 69, 90

J. P. Perdew and A. Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B*, 23(10):5048, 1981. 27

J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865, 1996a. 21, 68, 69, 76, 104

J. P. Perdew, K. Burke, and Y. Wang. Generalized gradient approximation for the exchange-correlation hole of a many-electron system. *Phys. Rev. B*, 54:16533, 1996b. 21

J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke. Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett.*, 100:136406, 2008. 21, 90, 104

J. P. Perdew, A. Ruzsinszky, G. I. Csonka, L. A. Constantin, and J. Sun. Workhorse semilocal density functional for condensed matter physics and quantum chemistry. *Phys. Rev. Lett.*, 103:026403, 2009. 22, 76, 104

J. P. Perdew, A. Ruzsinszky, J. M. Tao, V. N. Staroverov, G. E. Scuseria, and G. I. Csonka. Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *J. Chem. Phys.*, 123(6):062201, 2005. 20

Vivien Petzold. New density functionals with error estimation applied to atomic-scale systems. *PhD thesis at Technical University of Denmark*, October 2010. 64, 65, 66, 80

Vivien Petzold, Thomas Bligaard, and KW Jacobsen. Construction of new electronic density functionals with error estimation through fitting. *Topics in Catalysis*, 2012. URL `http://link.springer.com/article/10.1007/s11244-012-9801-7`. 40, 64, 66, 67, 68, 77

R. Podeszwa, K. Patkowski, and K. Szalewicz. Improved interaction energy benchmarks for dimers of biological relevance. *Phys. Chem. Chem. Phys.*, 12:5974, 2010. 33

J. Rezac, K. E. Riley, and P. Hobza. S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures. *J. Chem. Theory Comput.*, 7(8):2427, 2011. 33

G. Román-Pérez and J. M. Soler. Efficient implementation of a van der Waals density functional: Application to double-wall carbon nanotubes. *Phys. Rev. Lett.*, 103:096102, 2009. 24

Riccardo Sabatini, Tommaso Gorni, Stefano De Gironcoli, and Stefano de Gironcoli. Nonlocal van der Waals density functional made simple and efficient. *Physical Review B*, 87(4):041108, January 2013. ISSN 1098-0121. DOI: 10.1103/PhysRevB.87.041108. URL `http://link.aps.org/doi/10.1103/PhysRevB.87.041108`. 25

A. Salam. *Molecular Quantum Electrodynamics: Long-Range Intermolecular Interactions*. Wiley, first edition, 2009. 15

T. Sato and H. Nakai. Density functional method including weak interactions: Dispersion coefficients based on the local response approximation. *J. Chem. Phys.*, 131(22):224104, 2009. 23

L Schimka, J Harl, a Stroppa, a Grüneis, M Marsman, F Mittendorfer, and G Kresse. Accurate surface and adsorption energies from many-body perturbation theory. *Nature materials*, 9(9):741–4, September 2010. ISSN 1476-1122. DOI: 10.1038/nmat2806. URL `http://www.ncbi.nlm.nih.gov/pubmed/20657589`. 26

L. Schimka, J. Harl, and G. Kresse. Improved hybrid functional for solids: The HSEsol functional. *J. Chem. Phys.*, 134:024116, 2011. 35

C. D. Sherrill. Frontiers in electronic structure theory. *J. Chem. Phys.*, 132(11):110902, 2010. 33

J. C. Slater. Wavefunction in a periodic potential. *Phys. Rev*, 51:846, 1937. 29

V. N. Staroverov, G. E. Scuseria, J. M. Tao, and J. P. Perdew. Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *J. Chem. Phys.*, 119 (23):12129, 2003. 32

P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.*, 98:11623, 1994. 26, 76

A. Stroppa and G. Kresse. The shortcomings of semi-local and hybrid functionals: what we can learn from surface science studies. *New J. Phys.*, 10(6):063020, 2008.

J. Sun, M. Marsman, A. Ruszinszky, G. Kresse, and J. P. Perdew. Improved lattice constants, surface energies, and CO desorption energies from a semilocal density functional. *Phys. Rev. B*, 83:121410(R), 2011.

J. Sun, B. Xiao, and A. Ruzsinszky. Communication: Effect of the orbital-overlap dependence in the meta generalized gradient approximation. *J. Chem. Phys.*, 137(5):051101, 2012a. 22

Jianwei Sun, Bing Xiao, and Adrienn Ruzsinszky. Communication: Effect of the orbital-overlap dependence in the meta generalized gradient approximation. *The Journal of chemical physics*, 137(5): 051101, August 2012b. ISSN 1089-7690. DOI: 10.1063/1.4742312. URL http://www.ncbi.nlm.nih.gov/pubmed/22894323. 104

Jianwei Sun, Robin Haunschild, Bing Xiao, Ireneusz W Bulik, Gustavo E Scuseria, and John P Perdew. Semilocal and hybrid meta-generalized gradient approximations based on the understanding of the kinetic-energy-density dependence. *The Journal of chemical physics*, 138(4):044113, January 2013. ISSN 1089-7690. DOI: 10.1063/1.4789414. URL http://www.ncbi.nlm.nih.gov/pubmed/23387574. 22

T. Takatani, E. G. Hohenstein, M. Malagoli, M. S. Marshall, and C. D. Sherrill. Basis set consistent revision of the S22 test set of noncovalent interaction energies. *J. Chem. Phys.*, 132:144104, 2010. 33

J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria. Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.*, 91:146401, 2003. 22, 104

A. Tkatchenko and M. Scheffler. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.*, 102:073005, 2009. 23

Anja Toftelund. Error Estimation for Electronic Structure Calculations Supervised by. *Master Thesis, Technical University of Denmark*, November 2006. 63, 65

T. Van Voorhis and G. E. Scuseria. A novel form for the exchange-correlation energy functional. *J. Chem. Phys.*, 109(2):400, 1998. 22

O. A. Vydrov and T. Van Voorhis. Nonlocal van der Waals density functional made simple. *Phys. Rev. Lett.*, 103:063004, 2009. 25

O. A. Vydrov and T. Van Voorhis. Implementation and assessment of a simple nonlocal van der Waals density functional. *J. Chem. Phys.*, 132:164113, 2010a. 25

O. A. Vydrov and T. Van Voorhis. Comment on "Nonlocal van der Waals density functional made simple" Reply. *Phys. Rev. Lett.*, 104 (9):099304, 2010b. 25

O. A. Vydrov and T. Van Voorhis. Nonlocal van der Waals density functional: The simpler the better. *J. Chem. Phys.*, 133:244103, 2010. 25, 33

Oleg a Vydrov and Gustavo E Scuseria. Effect of the Perdew-Zunger self-interaction correction on the thermochemical performance of approximate density functionals. *The Journal of chemical physics*, 121(17):8187–93, November 2004. ISSN 0021-9606. DOI: 10.1063/1.1794633. URL http://www.ncbi.nlm.nih.gov/pubmed/15511137. 27

Jess Wellendorff, Keld T. Lundgaard, Andreas Mø gelhøj, Vivien Petzold, David D. Landis, Jens K. Nø rskov, Thomas Bligaard, and Karsten W. Jacobsen. Density functionals for surface science: Exchange-correlation model development with Bayesian error estimation. *Physical Review B*, 85(23):235149, June 2012. ISSN 1098-0121. DOI: 10.1103/PhysRevB.85.235149. URL http://link.aps.org/doi/10.1103/PhysRevB.85.235149. 32, 34, 35, 68, 75, 76, 104

Jun Yan, Jens S. Hummelshø j, and Jens K. Nø rskov. Formation energies of group I and II metal oxides using random phase approximation. *Physical Review B*, 87(7):075207, February 2013. ISSN 1098-0121. DOI: 10.1103/PhysRevB.87.075207. URL http://link.aps.org/doi/10.1103/PhysRevB.87.075207. 26

VJ Yohai. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15(2):642–656, 1987. URL http://www.jstor.org/stable/10.2307/2241331. 57

Y. Zhao and D. G. Truhlar. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.*, 125:194101, 2006. 22, 33

Y. Zhao and D. G. Truhlar. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Account.*, 120:215, 2008. 25, 33

J. Zheng, Y. Zhao, and D. G. Truhlar. Representative benchmark suites for barrier heights of diverse reaction types and assessment of electronic structure methods for thermochemical kinetics. *J. Chem. Theory Comput.*, 3:569, 2007. 32

J. Zheng, Y. Zhao, and D. G. Truhlar. The DBH24/08 Database and Its Use to Assess Electronic Structure Model Chemistries for Chemical Reaction Barrier Heights. *J. Chem. Theory Comput.*, 5(4):808, 2009. 32

Paper I

Density Functionals for Surface Science: Exchange-correlation Model Development with Bayesian Error Estimation

J. Wellendorff, K. T. Lundgaard, A. Møgelhøj, V. Petzold, D. D. Landis, J. K. Nørskov, T. Bligaard, and K. W. Jacobsen Physical Review B 85, 235149 (2012)

Physical Review B 85, 235149 (2012)

Density functionals for surface science: Exchange-correlation model development with Bayesian error estimation

Jess Wellendorff,[1,*] Keld T. Lundgaard,[1] Andreas Møgelhøj,[1,2] Vivien Petzold,[1] David D. Landis,[1] Jens K. Nørskov,[2,3] Thomas Bligaard,[2,3] and Karsten W. Jacobsen[1]

[1]*Center for Atomic-Scale Materials Design (CAMD), Department of Physics, Building 307, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark*

[2]*SUNCAT Center for Interface Science and Catalysis, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California 94025, USA*

[3]*Department of Chemical Engineering, Stanford University, Stanford, California 94305, USA*

(Received 20 March 2012; revised manuscript received 14 May 2012; published 27 June 2012)

A methodology for semiempirical density functional optimization, using regularization and cross-validation methods from machine learning, is developed. We demonstrate that such methods enable well-behaved exchange-correlation approximations in very flexible model spaces, thus avoiding the overfitting found when standard least-squares methods are applied to high-order polynomial expansions. A general-purpose density functional for surface science and catalysis studies should accurately describe bond breaking and formation in chemistry, solid state physics, and surface chemistry, and should preferably also include van der Waals dispersion interactions. Such a functional necessarily compromises between describing fundamentally different types of interactions, making transferability of the density functional approximation a key issue. We investigate this trade-off between describing the energetics of intramolecular and intermolecular, bulk solid, and surface chemical bonding, and the developed optimization method explicitly handles making the compromise based on the directions in model space favored by different materials properties. The approach is applied to designing the Bayesian error estimation functional with van der Waals correlation (BEEF–vdW), a semilocal approximation with an additional nonlocal correlation term. Furthermore, an ensemble of functionals around BEEF–vdW comes out naturally, offering an estimate of the computational error. An extensive assessment on a range of data sets validates the applicability of BEEF–vdW to studies in chemistry and condensed matter physics. Applications of the approximation and its Bayesian ensemble error estimate to two intricate surface science problems support this.

DOI: 10.1103/PhysRevB.85.235149 PACS number(s): 71.15.Mb, 31.15.eg, 68.43.−h

I. INTRODUCTION

Kohn-Sham density functional theory[1,2] (KS–DFT) is a widely celebrated method for electronic-structure calculations in physics, chemistry, and materials science.[3,4] Indeed, modern DFT methods have proven valuable for elucidating mechanisms and fundamental trends in enzymatic and heterogeneous catalysis,[5–13] and computational design of chemically active materials is now within reach.[14–17] Successful use of DFT often relies on accurate but computationally tractable approximations to the exact density functional for the exchange-correlation (XC) energy. The generalized gradient approximation (GGA) is very popular due to a high accuracy-to-cost ratio for many applications, but suffers from a range of shortcomings. Thus, common GGA functionals are well suited for computing many important quantities in chemistry and condensed matter physics, but appear to be fundamentally unable to accurately describe the physics and chemistry of a surface at the same time.[18] Moreover, van der Waals (vdW) dispersion interactions are not accounted for by GGAs,[19] and spurious self-interaction errors can be significant.[20–22] The interest in applying DFT to more and increasingly complex problems in materials science is not likely to decrease in the years to come. Much effort is therefore devoted to improve on current density functional approximations.

The five-rung "Jacob's ladder" of Perdew[23] represents a classification of the most popular density functional methods. Each rung adds new ingredients to the density functional approximation (DFA), and so should enable better approximations, but also adds to the computational cost. In order of increasing complexity, the ladder consists of the local spin-density approximation[1] (LDA), GGA, meta-GGA (MGGA), hyper-GGA, and finally the generalized random phase approximation (RPA). The LDA uses only the local density as input, while rungs 2 and 3 introduce semilocal dependence of the density (GGA) and the KS orbitals (MGGA).[24] Hyper-GGAs introduce nonlocal dependence of the occupied KS orbitals in the exact exchange energy density, and fifth-rung approximations calculate correlation energies from the unoccupied KS orbitals. The latter is computationally heavy, but RPA-type methods are the only DFAs in this five-rung hierarchy that can possibly account for vdW dispersion between nonoverlapped densities.[24]

The failure of lower-rung DFAs in capturing dispersion forces has spurred substantial developments in recent years.[19] Such interactions are spatially nonlocal in nature, and several different approaches to add "vdW terms" to lower-rung DFAs now exist.[25–28] The vdW–DF nonlocal correlation[25] is a particularly promising development in this field. It is a fully nonlocal functional of the ground-state density, and has proven valuable in a wide range of sparse matter studies.[29] However, the vdW–DF and vdW–DF2 (Ref. 30) methods yield much too soft transition-metal crystal lattices,[31,32] and the correct choice of GGA exchange functional to use in vdW–DF type calculations is currently investigated.[30,32–34] One approach to choosing GGA exchange is comparison to Hartree-Fock exchange[35,36] and consideration of the behavior of the exchange functional in the limit of large density gradients.[35] Where does the vdW–DF approximation belong in a hierarchy such as Jacob's ladder? In terms of computational complexity, the method contains fully nonlocal density-density information without explicit use of the KS orbitals. From this point of view, it should fit between rungs 3 and 4, and we assign it here to rung 3.5. Note that nonlocal exchange approximations, designed to partially mimic exact exchange at a reduced computational cost, have recently been proposed[37,38] as belonging to a rung 3.5.

Put in simple terms, two paradigms for developing density functionals are dominant: that of constraint satisfaction by reduction[24] and that of fitting to empirical data.[39–42] Both have contributed greatly to the success of DFT. Reductionists impose constraints based on analytic properties of the exact density functional, and strive for nonempirical functionals that fulfill as many constraints as possible on each rung of Jacob's ladder. Empirically oriented DFA developers use experimental or high-level theoretical training data to optimize the DFA description of one or more materials properties. Reduction is arguably the most systematic approach to density functional development, and has had a significant impact on the field of KS–DFT. However, choices are often made as to what types of physics and chemistry the DFA should describe well.[43,44] The empirical approach is fundamentally a matter of explicitly making these choices, and parametrize an XC model to suit personal preferences for computational performance. This makes overfitting the training data and transferability of the optimized DFA to systems and materials properties not contained in the training data a central issue.[24]

The risk of overfitting was realized early on by Becke and others.[40,45] Using polynomial expansions of GGA exchange and correlation in least-squares-fitting procedures, polynomial orders above four were found to yield increasingly oscillatory and unphysical XC functionals, that is, "a transition to mathematical nonsense."[45] Nevertheless, semiempirical DFAs containing many parameters have been constructed[42,46,47] with little attention to the overfitting issue. Transferability of a DFA parametrization depends not only on the degree of overfitting to a single set of molecular or condensed matter properties, but also on how many physically different properties the approximate model was trained on. Optimizing XC parametrizations to several different properties naturally leads to a "competition" between data sets in determining the model solution, i.e., an XC model compromise. Implicitly acknowledging this, each data set is often assigned more or less arbitrary weights.[46,47] In our view, such an approach is not guaranteed to yield the optimum model compromise.

In this study, we apply machine-learning methods to avoid the above-mentioned pitfalls of semiempirical density functional development. Regularization of a very flexible polynomial GGA exchange expansion is at the heart of the developed approach. We furthermore investigate the characteristics of XC model compromises in a GGA + vdW model space, and formulate and apply an explicit principle for how an XC model trade-off should be composed. Using several training data sets of quantities representing chemistry, solid state physics, surface chemistry, and vdW dominated interactions, the Bayesian error estimation functional with van der Waals (BEEF–vdW) exchange-correlation model is generated. The three most important aspects of semiempirical DFA design are thus considered in detail: data sets, model space, and model selection. The developed approach furthermore leads to an ensemble of functionals around the optimum one, allowing an estimate of the computational error to be calculated. Lastly, BEEF–vdW is evaluated on systems and properties partly not in the training sets, and is also applied in two small surface science studies: calculating potential-energy curves for graphene adsorption on the Ni(111) surface, and investigation of the correlation between theoretical chemisorption energies and theoretical surface energies of the substrate.

II. DATA SETS

Several sets of energetic and structural data describing bonding in chemical and condensed matter systems are used throughout this study. These data sets are either adapted from literature or compiled here from published works, and are briefly presented in the following. Additional information is found in the Appendix.

(a) *Molecular formation energies*. The G3/99 (Ref. 48) molecular formation enthalpies of Curtiss and co-workers represent intramolecular bond energetics. Experimental room-temperature heats of formation are extrapolated to 0 K, yielding 223 electronic-only static-nuclei formation energies. The G2/97 (Ref. 49) set of 148 formation energies is a subset of G3/99.

(b) *Molecular reaction energies*. Molecular formation energies lend themselves well to compilation of gas-phase reaction energies. The RE42 data set of 42 reaction energies involves 45 different molecules from G2/97.

(c) *Molecular reaction barriers*. The DBH24/08 (Ref. 50) set of Zheng *et al.*, comprising 12 forward and 12 backward benchmark barriers, is chosen to represent gas-phase reaction barriers.

(d) *Noncovalent interactions*. The S22 (Ref. 51) and S22x5 (Ref. 52) sets of intermolecular interaction energies of noncovalently bonded complexes calculated at the coupled-cluster level with single, double, and perturbative triple excitations [CCSD(T)] were compiled by Hobza and co-workers. Particularly, the S22 set has become popular for assessment[34,53–58] and parametrization[30,33,47,54,59,60] of density functional methods for vdW–type interactions. The S22x5 set consists of potential-energy curves (PECs) for each S22 complex, with interaction energies at relative interaction distances d of 0.9, 1.0, 1.2, 1.5, and 2.0 as compared to S22, totaling 110 data points. For convenience, this study divides S22x5 into five subsets according to interaction distance, e.g., "S22x5-0.9."
The accuracy of the original S22 and S22x5 energies have certain deficiencies, so the revised S22x5-1.0 energies of Takatani *et al.*[61] are used instead. The remaining (nonequilibrium) data points on each CCSD(T) PEC are correspondingly corrected according to the difference between original and revised S22x5-1.0 energies, as elaborated on in the Appendix.

(e) *Solid state properties*. Three sets of 0-K experimental solid state data are used, here denoted Sol34Ec, Sol27LC, and Sol27Ec. The first comprises cohesive energies of 34 period 2–6 bulk solids in fcc, bcc, diamond, and hcp lattices. Zero-point phonon effects have not been corrected for. Conversely, the Sol27LC and Sol27Ec sets contain lattice constants and cohesive energies, respectively, of 27 cubic lattices, both corrected for zero-point vibrational contributions.

(f) *Chemisorption on solid surfaces*. The CE17 and CE27 data sets comprise experimental reaction energies for chemisorption of simple molecules on the (111), (100), and (0001) facets of late transition-metal surfaces at low coverage. The CE17 set is a subset of CE27.

III. COMPUTATIONAL DETAILS

Self-consistent density functional calculations are performed using GPAW,[62–64] a real-space grid implementation of the projector augmented-wave method.[65] The ASE (Refs. 64 and 66) package provides a convenient interface to GPAW. Grid-point spacings of 0.16 Å are employed for high-quality computations of simple properties such as molecular binding energies. Properties of bulk solids are calculated using somewhat denser grids with a spacing of 0.13 Å. Real-space structure relaxation is applied to the G3/99 molecules and CE27 chemisorption systems with 0.05 eV/Å as the criterion of maximum force on each relaxing atom. Molecular and single-atomic systems are placed in a box with at least 7 Å vacuum to the box boundaries, except for the S22x5 complexes for which the vacuum width is 10 Å. Further details on the computational procedure employed are found in the Appendix.

IV. MODEL SPACE

The GGA exchange energy density $\varepsilon_x^{\mathrm{GGA}}(n,\nabla n)$ is conveniently expressed in terms of the exchange energy density of the uniform electron gas $\varepsilon_x^{\mathrm{UEG}}(n)$ and an exchange enhancement factor $F_x(s)$, depending on the local density as well as its gradient through the reduced density gradient s,

$$
\begin{aligned}
s &= \frac{|\nabla n|}{2k_F n}, \quad 0 \leqslant s \leqslant \infty,\\
\varepsilon_x^{\mathrm{GGA}}(n,\nabla n) &= \varepsilon_x^{\mathrm{UEG}}(n) F_x[s(n,\nabla n)], \qquad (1)\\
E^{\mathrm{GGA\text{-}x}}[n,\nabla n] &= \int \varepsilon_x^{\mathrm{UEG}}(n) F_x[s(n,\nabla n)] d\mathbf{r},
\end{aligned}
$$

where $n = n(\mathbf{r})$, $k_F = (3\pi^2 n)^{1/3}$ is the Fermi wave vector of the UEG, and $E^{\mathrm{GGA-x}}$ is the semilocal GGA exchange energy.

In this study, a highly general exchange model space is obtained by expanding the GGA exchange enhancement factor in a basis of M_x Legendre polynomials $B_m[t(s)]$ of orders 0 to $M_x - 1$ in a transformed reduced density gradient, denoted $t(s)$:

$$
\begin{aligned}
t(s) &= \frac{2s^2}{4+s^2} - 1, \quad -1 \leqslant t \leqslant 1\\
F_x^{\mathrm{GGA}}(s) &= \sum_m a_m B_m[t(s)],\\
E^{\mathrm{GGA\text{-}x}}[n,\nabla n] &= \sum_m a_m \int \varepsilon_x^{\mathrm{UEG}}(n) B_m[t(s)] d\mathbf{r} \qquad (2)\\
&= \sum_m a_m E_m^{\mathrm{GGA\text{-}x}}[n,\nabla n],
\end{aligned}
$$

where a_m are expansion coefficients, and $E_m^{\mathrm{GGA\text{-}x}}$ is the exchange energy corresponding to the Legendre basis function B_m. The polynomial basis is constructed such that the boundary limits $t = [-1,1]$ are zero for all $m > 1$ basis functions. Therefore, these limits are determined by the order 0 and 1 basis functions only.

Semilocal approximations to electron correlation effects beyond GGA exchange are not easily cast in terms of a single variable, such as s. The correlation model space is chosen to be a linear combination of the Perdew-Burke-Ernzerhof (PBE) (Ref. 67) semilocal correlation functional, purely local Perdew-Wang[68] LDA correlation, and vdW–DF2 (Ref. 30) type nonlocal correlation. The latter is calculated from a double integral over a nonlocal interaction kernel $\phi(\mathbf{r},\mathbf{r}')$,

$$
E^{\mathrm{nl\text{-}c}}[n] = \frac{1}{2} \int n(\mathbf{r})\phi(\mathbf{r},\mathbf{r}')n(\mathbf{r}') d\mathbf{r}\, d\mathbf{r}', \qquad (3)
$$

which is evaluated using the fast Fourier transformation method of Román-Pérez and Soler,[69] implemented in GPAW as described in Ref. 70.

In total, the XC model space consequently consists of GGA exchange expanded in Legendre polynomials as well as local, semilocal, and nonlocal correlation,

$$
\begin{aligned}
E_{xc} = &\sum_{m=0}^{M_x-1} a_m E_m^{\mathrm{GGA\text{-}x}} + \alpha_c E^{\mathrm{LDA\text{-}c}}\\
&+ (1-\alpha_c) E^{\mathrm{PBE\text{-}c}} + E^{\mathrm{nl\text{-}c}}, \qquad (4)
\end{aligned}
$$

where $M_x = 30$, and the total number of parameters is $M = M_x + 1 = 31$.

None of the commonly imposed constraints on GGA exchange are invoked, e.g., the LDA limit of $F_x(s)$ and recovery of the correct gradient expansion for slowly varying densities, nor the Lieb-Oxford (LO) bound[71,72] for large electron density gradients. However, as seen from Eq. (4), the sum of LDA and PBE correlation is constrained to unity.

V. MODEL SELECTION

Choices are made when developing a semiempirical density functional. These are both explicit and implicit choices pertaining to what the functional is to be designed for, that is, for the selection of an optimum exchange-correlation model that captures the materials properties of main interest when applying the approximation. This study aims to explicate the choices, and to develop a set of principles for the model selection process. These principles are used to guide the inevitable compromise between how well significantly different quantities in chemistry and condensed matter physics are reproduced by an incomplete XC model space. Development of an XC functional is in this approach divided into two steps. First an individual model selection for a number of data sets is carried out, and subsequently a simultaneous model selection is made, compromising between the individual fits.

A. Individual materials properties

1. *Regularizing linear models*

Model training is formulated in terms of finding the expansion coefficient vector that minimizes a cost function without overfitting the data. This may be viewed as determining the optimum trade-off between bias and variance of the model.[73] The cost function contains two terms: a squared error term and a regularization term. One simple regularization suitable for varying the bias-variance ratio is one that "penalizes" the cost function for model solutions that differ from a suitably chosen prior solution. This effectively removes sloppy[74] eigenmodes of the cost function by adding curvature to all modes, and thereby limits the effective number of parameters in the model solution. As the regularization strength is continuously decreased from infinity towards zero, the model parameters that minimize the cost function are allowed to differ increasingly from the prior solution. In a sufficiently large model space, the solution that reproduces the data best without overfitting is in general found for intermediate regularization strength. A slightly more elaborate regularization is used in this study, as outlined later on.

Finding the optimum model is then a matter of determining the optimum regularization strength. This may be done by minimizing the estimated prediction error (EPE) for varying regularization strength. The EPE provides a statistical estimate of the validity of a model outside the space of its training data, and can be obtained by a large variety of resampling methods. We obtain it using bootstrap resampling.[75] Even though common error quantities, such as the standard deviation (STD), will in general decrease for regularization strengths smaller than that which minimizes the EPE, the corresponding model solutions are likely to be increasingly overfitted. Minimizing the EPE and not the STD is therefore preferred for determining well-behaved XC functionals.

2. *Details of the procedure*

The standard Tikhonov regularization method[73] is chosen to control overfitting. A cost function for the ith data set is therefore defined as

$$C_i(\mathbf{a}) = (\mathbf{X}_i\mathbf{a} - \mathbf{y}_i)^2 + \omega^2\Gamma^2(\mathbf{a} - \mathbf{a}_p)^2, \qquad (5)$$

where $\mathbf{X}_i$ is a data matrix, $\mathbf{a}$ the coefficient vector, $\mathbf{y}_i$ a target vector of training data, ω^2 the regularization strength, Γ is denoted the Tikhonov matrix, and the prior vector $\mathbf{a}_p$ is the origo for regularization, i.e., the model solution for $\omega^2 \to \infty$ and thus the model space reference point for regularization.

In accordance with Eq. (4), the data matrix consists of XC contributions to a materials property for each system in the ith data set from the M basis functions. These are evaluated non-self-consistently on revised PBE (RPBE) (Ref. 76) densities. The target vector contains the target XC contribution to each quantity in the set. The Tikhonov matrix is defined from a smoothness criterion on the basis functions. The exchange part of Γ is the overlap of the second derivative of the exchange basis functions with respect to the transformed reduced density gradient

$$\mathbf{\Gamma}^2_{ij} = \int_{-1}^{1} \frac{d^2B_i(t)}{dt^2}\frac{d^2B_j(t)}{dt^2}dt. \qquad (6)$$

Defined this way, the Tikhonov matrix directly penalizes the integrated squared second derivative of the exchange fit for finite regularization strength. This can be understood as penalizing a measure of nonsmoothness of the fitted exchange enhancement factor. In effect, the $\mathbf{\Gamma}$ matrix scales the regularization strength acting on each exchange basis function, such that higher-order basis functions are suppressed when minimizing the cost function. This leads to a model selection preference for solution vectors with small coefficients for higher-order polynomials, unless they are essential for obtaining a satisfactory fit. Physically, it is very reasonable to require $F_x(s)$ to be a smooth and preferably injective function of s, and significantly nonsmooth exchange solutions have been shown to degrade transferability of fitted exchange functionals to systems outside the training data.[77] The correlation part of Γ has one in the diagonal and zeros in the off-diagonal elements. Since $\mathbf{\Gamma}$ acts in the transformed $t(s)$ space, the transformation in Eq. (2) causes the regularization penalty on exchange to be strongest in the large-s regime, where information from the data matrix about the optimum behavior of $F_x(s)$ is expected to be scarce.[76,78]

In order to minimize the cost function in Eq. (5), it is transformed by $\mathbf{\Gamma}^{-1}$. Ones are therefore inserted in the first two diagonal elements of $\mathbf{\Gamma}$ to avoid numerical issues. The solution vector $\mathbf{a}_i$ that minimizes C_i is written as

$$\mathbf{a}_i = \mathbf{\Gamma}^{-1}\left(\mathbf{X}_i'^T\mathbf{X}_i' + \mathbf{L}^2\omega_i^2\right)^{-1}\left(\mathbf{X}_i'^T\mathbf{y}_i + \omega^2\mathbf{L}^2\mathbf{a}_p'\right), \qquad (7)$$

where $\mathbf{X}_i' = \mathbf{X}_i\mathbf{\Gamma}^{-1}$, $\mathbf{a}_p' = \mathbf{\Gamma}\mathbf{a}_p$, and $\mathbf{L}^2$ is the identity matrix with zeros in the first two diagonal elements. Singular value decomposition of $\mathbf{X}_i'^T\mathbf{X}_i'$ is used to calculate the inverse matrix. The LDA and PBE correlation coefficients in the XC model are constrained to be between 0 and 1, implying $\alpha_c \in [0,1]$ for the correlation coefficient in Eq. (4). In the cases that this is not automatically fulfilled, it is enforced by recalculating the solution while fixing α_c to the nearest bound of the initial solution.

The exchange part of the prior vector is chosen as the linear combination of the order 0 and 1 polynomial basis functions that fulfills the LDA limit at $s = 0$ and the LO bound for $s \to \infty$. With the exchange basis transformation in Eq. (2), the prior for exchange is quite close to the PBE exchange enhancement factor. For $\omega^2 \to \infty$, we therefore nearly recover PBE exchange, while lower regularization strengths allow increasingly nonsmooth variations away from this prior solution. The optimum model is expected to include at least some semilocal correlation,[31] so the origo of correlation is $\alpha_c = 0.75$.

As previously mentioned, the optimum regularization is found by minimizing the estimated prediction error for varying ω^2. Bootstrap resampling of the data matrix with the .632 estimator[75,79] is used. It is defined as

$$\text{EPE}_{.632} = \sqrt{0.368 \cdot \widehat{\text{err}} + 0.632 \cdot \widehat{\text{Err}}}, \qquad (8)$$

where $\widehat{\text{err}}$ is the variance between the target data and the prediction by the optimal solution $\mathbf{a}_i$, and $\widehat{\text{Err}}$ measures the variance on samples of data to which solutions were not fitted in the resampling. Both are determined as a function of ω^2, and $\widehat{\text{Err}}$ is given by

$$\widehat{\text{Err}} = \frac{1}{N_\mu}\sum_\mu \frac{1}{N_{s|\mu\notin s}}\sum_{s|\mu\notin s}(\mathbf{x}_\mu\mathbf{b}_s - y_\mu)^2, \qquad (9)$$

where μ is an entry in the data set, N_μ the number of data points, s a bootstrap sample of N_μ data points, and $N_{s|\mu\notin s}$ the number of samples not containing μ. The parentheses calculate the difference between the prediction $\mathbf{x}_\mu\mathbf{b}_s$ of the data point μ by the best-fit coefficient vector $\mathbf{b}_s$ and the μth target value y_μ.

TABLE I. Model selection results of individually training the XC model of Eq. (4) to 10 different data sets. M_{eff} is the effective number of parameters in a model [see Eq. (21)]. The $s = 0$ and $s \to \infty$ limits of the obtained exchange enhancement factors are also shown. MSD, MAD, and STD are mean signed, mean absolute, and standard deviation, respectively, all in meV. Note that these are non-self-consistent results.

	α_c	M_{eff}	$F_x(0)$	$F_x(\infty)$	MSD	MAD	STD
CE17	0.90	4.7	0.97	2.15	−10	96	116
RE42	1.00	4.2	1.06	1.21	19	168	207
DBH24/08	0.00	3.7	1.14	3.14	1	116	142
G2/97	0.27	7.2	1.10	2.53	−13	109	149
Sol34Ec	0.00	7.7	0.97	1.25	−4	168	208
S22x5-0.9	0.81	3.2	0.96	1.68	0	9	11
S22x5-1.0	0.82	3.1	0.98	1.87	0	8	10
S22x5-1.2	0.40	5.7	1.04	2.38	0	4	6
S22x5-1.5	0.85	4.0	1.02	1.91	−1	3	4
S22x5-2.0	1.00	3.3	0.95	1.37	2	3	3

The best-fit solution is found by minimizing the cost function with the data in sample s only.

In the bootstrap resampling procedure, 500 randomly generated data samples are selected independently for each ω^2. The regularization strength that minimizes the .632 estimator is found by a smooth fitting of the slightly scattered estimator plot near the minimum. To properly regularize the S22x5 subsets with long interaction distances, a condition $F_x(s = \infty) \geqslant 1$ is enforced.

3. *Individually trained XC models*

Table I and Fig. 1 show details and statistics for the optimized XC models obtained when the procedure outlined above is applied to molecular, solid state, surface chemical, and vdW dominated energetics. Each model is therefore trained on a single materials property only, and their features differ significantly.

The DBH24/08 set appears to favor GGA exchange that substantially violates the LDA limit [$F_x(0) = 1.14$] along with inclusion of full PBE correlation ($\alpha_c = 0$; no LDA correlation). The model furthermore overshoots the LO bound $F_x^{\mathrm{LO}} = 1.804$ significantly [$F_x(\infty) = 3.14$]. The XC model optimized to the G2/97 set shows similar trends with respect to GGA exchange and PBE correlation, but is less extreme. In the other end of the spectrum is the model optimized to the Sol34Ec cohesive energies. These favor GGA exchange starting out slightly below $F_x = 1$, then reaching a maximum at $s \approx 2$, and finally declining slowly towards $F_x = 1.25$. Best agreement with experimental cohesive energies is found with full PBE correlation in addition to nonlocal correlation. The occurrence of a maximum in the exchange enhancement factor should, however, not be overemphasized. It has been shown[76,78] that only small GGA exchange contributions to chemical and solid state binding energetics can be attributed to reduced density gradients above 2.5. In the region of large s, where the smoothness criterion on exchange is strongly enforced, the regularization term in the cost function [Eq. (5)] will therefore be dominant in determining the solution for such systems. The regularization may therefore well determine the behavior of $F_x(s)$ for large density gradients.

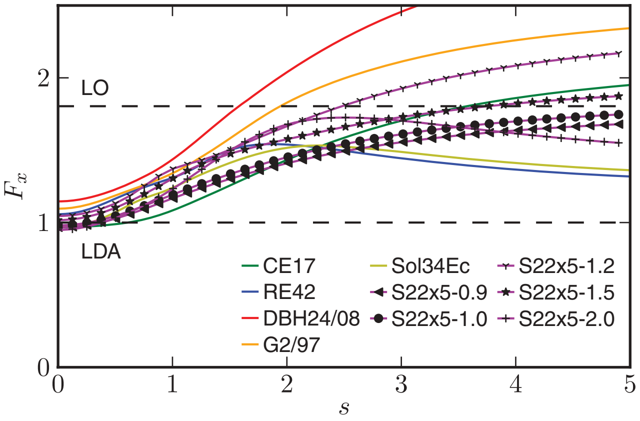


FIG. 1. (Color online) Exchange enhancement factors of the individually trained XC models listed in Table I.

For the remaining data sets in Table I, the optimized XC models appear reasonable, with all exchange enhancement factors starting out near the LDA limit. It is illustrative to investigate how the XC models perform for data sets on which they were not trained. The standard deviation is a natural measure of performance. Defining the relative standard deviation rSTD on some data set with some XC model, as the STD obtained by that model divided by the STD of the model that was fitted to that data set, rSTD is a measure of transferability. Figure 2 shows a color map of the rSTD for all 10 training data sets with all 10 trained models. The diagonal from bottom left to top right is, by definition, ones. In a background of blue and yellow-green squares, the map features two distinct areas of mostly reddish squares. To the far right, the S22x5-2.0 model yields rSTD > 5 for all other sets than DBH24/08, and rSTD ≈ 28 for S22x5-0.9. Furthermore, a 5 × 4 square in the top left corner illustrates that XC models trained on chemical or solid state data sets perform significantly worse on vdW–type energetics than models fitted to the latter. It is also interesting to see that the S22x5-2.0 rSTDs are largely unaffected by changing XC models. With little or no density-density overlap between many of the S22x5-2.0 complexes, the constant nonlocal correlation in all 10 models is likely the main XC contribution to intermolecular binding.

In summary, the deviation statistics in Table I illustrate that the XC model space considered here most certainly spans the

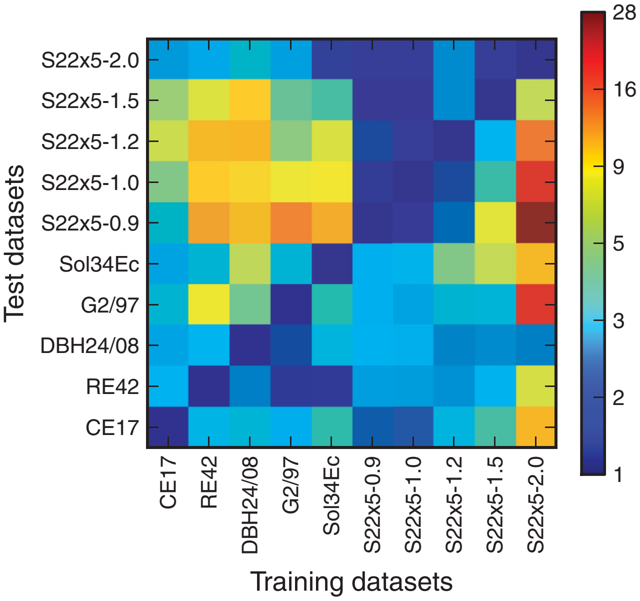


FIG. 2. (Color online) Color map of the relative standard deviations obtained when non-self-consistently applying the XC models found individually for each training data set, listed on the abscissa, to all 10 training data sets along the ordinate.

model degrees of freedom necessary to obtain well-performing density functionals with smooth exchange enhancement factors and sound correlation components. However, a high degree of transferability between the data sets should not be expected for several of the models.

B. Several materials properties

Fundamentally, a compromise has to be made between how well different materials properties are reproduced by the same semiempirical density functional. This is expressed as a compromise between how well the functional quantitatively performs on different training data sets. What the compromise should be can only be determined by the target applications of the functional, and one challenge is to make this choice as explicit as possible. This section presents one route towards a methodology for optimizing an XC model to simultaneously describe several different materials properties. First, the nature of the model compromise is illustrated for the case of simultaneously fitting two data sets using a summed cost function with varying weights on the two sets. However, in the end, a product cost function is found more convenient for determining the optimum weights according to the directions in model space favored by different data sets.

1. *Model compromise*

Consider first the problem of simultaneously fitting two data sets, and let the model compromise be described through the total cost function, given as the sum of the two individual cost functions:

$$\Lambda(\mathbf{a}) = \mathcal{W}_1 C_1(\mathbf{a}) + \mathcal{W}_2 C_2(\mathbf{a}), \tag{10}$$

where $\mathcal{W}_i$ is a weight on data set i. The coefficient vector solution $\mathbf{b}$ that minimizes $\Lambda(\mathbf{a})$ is found by setting the derivative to zero: Since the summed cost function is quadratic in $\mathbf{a}$, as the individual cost functions C_i are, it may be expressed in terms of the individual solutions $\mathbf{a}_i$ as

$$\Lambda(\mathbf{a}) = \sum_{i=1,2} \mathcal{W}_i \left(C_i^0 + \frac{1}{2}(\mathbf{a}-\mathbf{a}_i)^T \mathbf{H}_i (\mathbf{a}-\mathbf{a}_i) \right), \tag{11}$$

where $C_i^0 = C_i(\mathbf{a}_i)$ is the minimized cost of data set i, and $\mathbf{H}_i$ is the Hessian of $C_i(\mathbf{a})$. The minimizing solution $\mathbf{b}$ is thus found from the individual solutions $\mathbf{a}_i$ as

$$\mathbf{b} = \left(\sum_{i=1,2} \mathcal{W}_i \mathbf{H}_i \right)^{-1} \left(\sum_{i=1,2} \mathcal{W}_i \mathbf{H}_i \mathbf{a}_i \right). \tag{12}$$

However, a principle for guiding the choice of weights is needed.

Let us consider establishing a compromise based on explicit principles. The regularized cost functions for each training data set $C_i(\mathbf{a})$ contain information of the costs associated with deviating from the individually found model solutions $\mathbf{a}_i$ along all directions in model space. The individual costs all increase when moving away from $\mathbf{a}_i$ due to deterioration of the fits, increased overfitting, or a combination of both. Define now the relative cost for each data set, rCost[i], as the individual cost for set i evaluated at the compromising solution $\mathbf{b}$ relative to the individual cost at $\mathbf{a}_i$, hence

$$\text{rCost}[\,i\,] = \frac{C_i(\mathbf{b})}{C_i(\mathbf{a}_i)} = \frac{C_i(\mathbf{b})}{C_i^0} \geqslant 1. \tag{13}$$

Thus defined, the relative cost for each training data set is a simple measure of how unfavorable it is for each data set to be fitted by the compromising solution $\mathbf{b}$ instead of the individual solutions $\mathbf{a}_i$.

The main panel of Fig. 3 illustrates XC model compromises between the G2/97 and S22x5-1.0 data sets. The curve maps out the relative costs on both data sets obtained from model solutions $\mathbf{b}$ when systematically varying the weights in $\Lambda(\mathbf{a})$. The weight fraction f is introduced (see caption of Fig. 3). A wide range of poorly compromising models can obviously be produced, sacrificing a lot of relative cost on one set while gaining little on the other. However, if both materials properties represented by the two data sets are considered important, the optimum weightening is somewhere midway between the asymptotic extrema.

The inset in Fig. 3 shows how the product of the relative costs varies with f. To the right along the abscissa, where the fraction increasingly favors the G2/97 set, the rCost product increases rapidly. To the left, the increase is much smaller, but a minimum is located in-between. At least one intermediate minimum is always present since the slopes in the two asymptotic regions are $-\infty$ and 0, respectively. This property is induced by the variational property around the two original minima of the individual cost functions. Similar conclusions apply to any combination of two or more data sets that do not favor the same directions in the incomplete model space.

We find in general that the condition of minimizing the product of relative costs is well suited for choosing cost function weights for arbitrary numbers of training data sets, if the aim is a general-purpose model. This condition, which

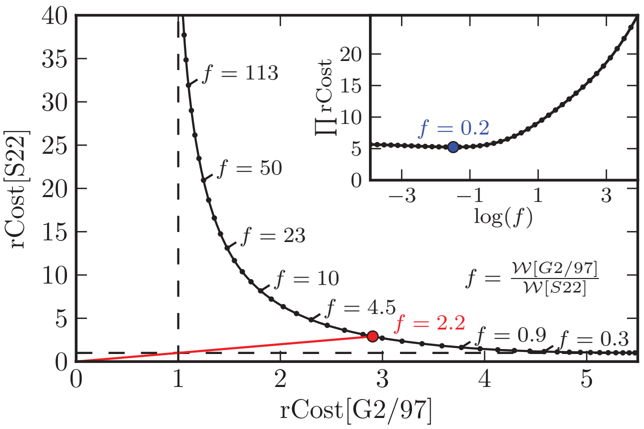


FIG. 3. (Color online) Main panel: XC model compromises between the G2/97 and S22x5-1.0 data sets illustrated in terms of relative costs (rCost) for both data sets when the weight fraction $f = \mathcal{W}[\text{G2/97}]/\mathcal{W}[\text{S22x5-1.0}]$ is varied and the summed cost function Eq. (10) is minimized. A range of compromising solutions are obtained, many of which are essentially fitting one data set only (rCost $\approx$ 1) while sacrificing model performance on the other (rCost $\gg$ 1). A red dot marks the point of equal rCost. The fact that an XC model with rCost[G2/97] = rCost[S22x5-1.0] = 1 is not obtainable illustrates the necessity of a model compromise. Inset: The product of relative costs display a minimum (blue dot) for a certain weight fraction.

is identical to minimizing the product of costs, is applied henceforth.

2. *Product cost function*

A product cost function for arbitrary numbers of training data sets is here defined, such that the minimizing solution **c** yields a desired minimum of the product of costs. The cost function is written as

$$\Phi(\mathbf{a}) = \prod_i C_i(\mathbf{a})^{w_i}, \tag{14}$$

where w_i is a constant weight, and C_i is again an individual cost function. The constant weight is an important feature of $\Phi(\mathbf{a})$ since it allows inclusion of training data sets which are perceived significantly less important than others. It is thus chosen from personal preferences given the purpose of the functional, and we shall see that **c** minimizes the product of costs *given* this choice.

For the case of two data sets, the stationary point between the two individual solutions in model space is found by differentiating the logarithm of $\Phi(\mathbf{a})$ with respect to **a**, and solving

$$\sum_i \frac{w_i}{C_i}\frac{dC_i}{d\mathbf{a}} = 0. \tag{15}$$

Using the method outlined above, the model solution that minimizes $\Phi(\mathbf{a})$ is found in terms of the individual solutions as

$$\mathbf{c} = \left(\sum_i \frac{w_i}{C_i}\mathbf{H}_i\right)^{-1}\left(\sum_i \frac{w_i}{C_i}\mathbf{H}_i\mathbf{a}_i\right), \tag{16}$$

where $C_i = C_i(\mathbf{c})$, and w_i simply scales the individual costs. We see that this solution corresponds to letting $\mathcal{W}_i$ in Eq. (11) equal w_i/C_i. Thus, minimizing the product of costs has introduced a natural weight C_i^{-1}, while w_i still leave room for deliberately biasing the model solution.

From here on, the product solution is therefore used to find the desired XC model solution: Since C_i is evaluated at **c**, the optimum solution is found iteratively, using C_i^{-1} as an iterator while searching for a converged minimum of the product cost function, given the constant weights w_i.[80]

3. *BEEF–vdW density functional*

The BEEF–vdW exchange-correlation functional was designed using the set of weights w listed in Table II. In principle, these should all equal one, however, correlations between some of the data sets have led us to lower the constant weight on some of them: Since the RE42 set is based on G2/97 molecules, the data in RE42 are correlated with some of the data in G2/97. Both weights were therefore lowered to 0.5. The same reasoning applies to the S22x5 subsets, where the same complexes are found in all the five sets, albeit at different interaction distances. A weight of $1/5 = 0.2$ on each S22x5 subset would therefore be natural, but for reasons of performance of the final functional, constant weights of 0.1 were chosen. The prior vector was the same for the combined functional as for the individual models.

The resulting model compromise is also tabulated in Table II, showing the effective data-set weight w/C, rCost, and rSTD for all data sets used in model training. It is clearly seen that especially the S22x5-0.9 interaction energies are hard to fit simultaneously with the other data sets within the XC model space employed here: The relative cost for the set is high, allowing the model to adapt mostly to the other data sets by lowering w/C for this set. This is furthermore reflected in the rSTD of 5.4, indicating that the BEEF–vdW performance on this data set is significantly worse than obtained in the individual fit to the S22x5-0.9 systems reported in Table I. Even so, the remaining S22x5 subsets appear to share XC

TABLE II. The BEEF–vdW model compromise. The effective weight in determining the XC model solution is w/C for each data set, as iteratively found from minimizing the product cost function [Eq. (14)]. The relative standard deviation (rSTD) is the ratio of the STD at the BEEF–vdW compromise to the STD at the regularized individual solution in Table I. The relative costs (rCost) are defined similarly, but includes regularization [see Eq. (13)].

	w	w/C	rCost	rSTD
CE17	1.0	1.80	1.7	1.3
RE42	0.5	0.62	2.5	1.8
DBH24/08	1.0	0.65	4.9	2.3
G2/97	0.5	0.62	2.6	1.6
Sol34Ec	1.0	0.43	7.5	2.8
S22x5-0.9	0.1	0.01	28.6	5.4
S22x5-1.0	0.1	0.04	9.1	2.9
S22x5-1.2	0.1	0.09	3.5	2.1
S22x5-1.5	0.1	0.08	4.1	2.1
S22x5-2.0	0.1	0.18	1.8	1.5

TABLE III. Expansion coefficients a_m for the BEEF–vdW Legendre exchange basis functions of order m. The correlation mixing parameter, α_c in Eq. (4), is 0.6001664769.

m	a_m	m	a_m
0	$1.516501714 \times 10^{0}$	15	$-8.018718848 \times 10^{-4}$
1	$4.413532099 \times 10^{-1}$	16	$-6.688078723 \times 10^{-4}$
2	$-9.182135241 \times 10^{-2}$	17	$1.030936331 \times 10^{-3}$
3	$-2.352754331 \times 10^{-2}$	18	$-3.673838660 \times 10^{-4}$
4	$3.418828455 \times 10^{-2}$	19	$-4.213635394 \times 10^{-4}$
5	$2.411870076 \times 10^{-3}$	20	$5.761607992 \times 10^{-4}$
6	$-1.416381352 \times 10^{-2}$	21	$-8.346503735 \times 10^{-5}$
7	$6.975895581 \times 10^{-4}$	22	$-4.458447585 \times 10^{-4}$
8	$9.859205137 \times 10^{-3}$	23	$4.601290092 \times 10^{-4}$
9	$-6.737855051 \times 10^{-3}$	24	$-5.231775398 \times 10^{-6}$
10	$-1.573330824 \times 10^{-3}$	25	$-4.239570471 \times 10^{-4}$
11	$5.036146253 \times 10^{-3}$	26	$3.750190679 \times 10^{-4}$
12	$-2.569472453 \times 10^{-3}$	27	$2.114938125 \times 10^{-5}$
13	$-9.874953976 \times 10^{-4}$	28	$-1.904911565 \times 10^{-4}$
14	$2.033722895 \times 10^{-3}$	29	$7.384362421 \times 10^{-5}$

model space with the data sets representing formation and rupture of interatomic bonds to a significantly greater extent. Thus, accurate description of the balance of strong and weak interactions in the S22x5-0.9 complexes is nearly incompatible with at least one of the other sets of materials properties, when demanding well-behaved exchange and correlation functionals in the present model space.

Table III lists the BEEF–vdW expansion coefficients. The correlation functional consists of 0.6 LDA, 0.4 PBE, and 1.0 nonlocal correlation. The qualitative shape of the BEEF–vdW exchange enhancement factor is shown in Fig. 4, with $s = 0$ and $s \to \infty$ limits of 1.034 and 1.870, respectively. Thus, BEEF–vdW exchange does not exactly obey the LDA limit for $s = 0$, but is 3.4% higher. The enhancement factor is above most GGA exchange functionals up to $s \approx 2.5$, from where it approaches the LO bound with a small overshoot in the infinite limit.

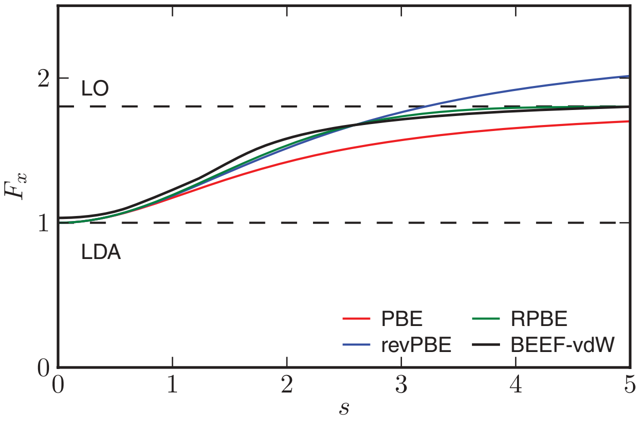


FIG. 4. (Color online) The BEEF–vdW exchange enhancement factor compared to those of a few standard GGA exchange functionals. The corresponding BEEF–vdW correlation functional is composed of 0.6 LDA, 0.4 PBE, and 1.0 nonlocal correlation.

The lack of exact fulfillment of the LDA limit for exchange indicates a conflict between this limit, the training data, and the employed preference for smooth exchange models. The G2/97 and DBH24/08 chemical data sets are found to give particular preference to exchange enhancement models with $F_x(0) \approx 1.1$, and enforcing $F_x(0) = 1.0$ for these sets leads to severely nonsmooth exchange solutions for $s \to 0$. Similar behavior was found in Ref. 77. Note that MGGA approximations are able to achieve exchange models with $F_x(0) \neq 1.0$ for densities different from the UEG, while still obeying the LDA limit for UEG-like densities. The BEEF–vdW F_x also has small "bump" at $s \approx 1.3$. This is not essential to the quality of the model and is not expected to harm its transferability. However, completely removing such features requires overly strong regularization.

VI. ENSEMBLE ERROR ESTIMATION

A normal DFT calculation does not provide any information about the uncertainty of the result from using an approximate XC functional. One method to obtain an estimate of the uncertainty is performing several calculations using different functionals, and observe the variations in the prediction of the quantity of interest. Another more systematic approach is to use an ensemble of functionals designed to provide an error estimate, as discussed in Ref. 81. This method is applied to the BEEF–vdW model, and the adaptation is briefly presented here.

Inspired by Bayesian statistics,[73] we define a probability distribution P for the model parameters $\mathbf{a}$ given the model θ and training data D:

$$P(\mathbf{a}|\theta D) \sim \exp[-C(\mathbf{a})/\tau], \tag{17}$$

where $C(\mathbf{a})$ is the cost function, and τ is a cost "temperature." Given the data D, a model perturbation $\delta\mathbf{a}$ has a certain probability associated with it, and this defines an ensemble of different XC functionals. The temperature is to be chosen such that the spread of the ensemble model predictions of the training data reproduces the errors observed when using BEEF–vdW self-consistently. This approach to constructing the probability distribution is closely related to the maximum entropy principle.[77,82]

The ensemble is defined through a Hessian scaled with the temperature. The Hessian is calculated directly from

$$\mathbf{H} = 2\sum_i^N \frac{w_i}{C_i(\mathbf{a}_p)} \mathbf{\Gamma}^{-1}\left(\mathbf{X}_i'^T\mathbf{X}_i' + \omega_i^2\mathbf{L}^2\right)\mathbf{\Gamma}^{-1^T}, \tag{18}$$

where the sum is over training data sets. The temperature is related to the effective number of parameters in the model, calculated from the effective regularization

$$\omega_{\text{eff}}^2 = \sum_i^N \frac{w_i}{C_i(\mathbf{c})}\omega_i^2, \tag{19}$$

where ω_i^2 are the regularization strengths for the individual data sets. Additionally, diagonalization of the combined square of the transformed data matrix

$$\Sigma' = \mathbf{V}^T\left(\sum_i^N \frac{w_i}{C_i(\mathbf{c})}\mathbf{X}_i'^T\mathbf{X}_i'\right)\mathbf{V}, \tag{20}$$

where Σ' contains the eigenvalues along the diagonal and $\mathbf{V}$ the eigenvectors, allows the effective number of parameters left in the model after regularization, $M_{\rm eff}$, to be computed as

$$M_{\rm eff} = \sum_{m}^{M} \frac{{\Sigma'_m}^2}{{\Sigma'_m}^2 + \omega_{\rm eff}^2 L_m^2}. \tag{21}$$

Since $M_{\rm eff} = 7.11$ in the BEEF–vdW model compromise, more than 75% of the initially 31 model degrees of freedom have been suppressed by regularization.

The temperature calculation is slightly modified from the method in Ref. 81 in order to construct an unbiased error estimation. This reflects that a larger error is expected when BEEF–vdW is applied to systems not included in the training data sets. The temperature is therefore calculated as

$$\tau = 2\frac{C(\mathbf{c})}{M_{\rm eff}} \cdot \frac{N_{\rm tot}}{N_{\rm tot} - M_{\rm eff}}, \tag{22}$$

where $N_{\rm tot}$ is the total number of systems in the training sets. The second term is close to unity since $N_{\rm tot} \gg M_{\rm eff}$. An ensemble matrix is now found as

$$\mathbf{\Omega}^{-1} = \tau\, \mathbf{H}^{-1}, \tag{23}$$

with eigenvalues $\mathbf{w}^2_{\Omega^{-1}}$ and eigenvectors $\mathbf{V}_{\Omega^{-1}}$.

Finally, using an ensemble of k vectors $\mathbf{v}_k$, each of length M with elements randomly drawn from a normal distribution of zero mean and variance one, the BEEF–vdW ensemble coefficient vectors $\mathbf{a}_k$ are calculated from

$$\mathbf{a}_k = \mathbf{V}_{\Omega^{-1}} \cdot \mathbf{1}\mathbf{w}_{\Omega^{-1}} \cdot \mathbf{v}_k. \tag{24}$$

The BEEF–vdW ensemble matrix is provided in the Supplemental Material.[83]

An illustration of the BEEF–vdW ensemble is shown in Fig. 5. For each data point in each data set, this ensemble may be applied non-self-consistently to BEEF–vdW electron densities. The standard deviation of the ensemble predictions of a quantity is then the ensemble estimate of the BEEF–vdW standard deviation on that quantity. The exchange enhancement ensemble expands after $s \approx 2$, where most of the chemistry and solid state physics have already happened.[76,78]

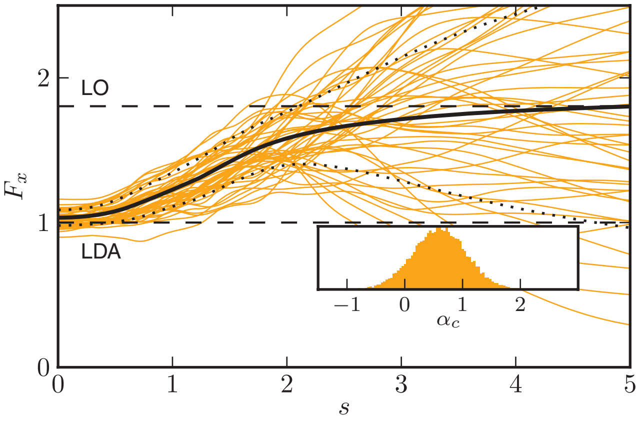


FIG. 5. (Color online) Bayesian ensemble of XC functionals around BEEF–vdW. Main panel: Black solid line is the BEEF–vdW exchange enhancement factor, while the orange lines depict $F_x(s)$ for 50 samples of the randomly generated ensemble. Dashed black lines mark the exchange model perturbations that yield DFT results ±1 standard deviation away from BEEF–vdW results. The inset shows a histogram of the distribution of correlation parameters in an ensemble containing 20 000 samples. The distribution is centered around $\alpha_c = 0.6$.

TABLE IV. Comparison of self-consistent BEEF–vdW standard deviations to those predicted by the ensemble of functionals around BEEF–vdW. All energies in meV.

	BEEF–vdW	Ensemble estimate
CE17	143	209
RE42	372	253
DBH24	331	144
G2/97	242	312
SolEc34	576	436
s22x5-0.9	171	197
s22x5-1.0	94	181
s22x5-1.2	36	137
s22x5-1.5	8	67
s22x5-2.0	5	18

The predictive performance of the ensemble has been evaluated using 20 000 ensemble functionals. In practice, however, a few thousand ensemble functionals suffice for well-converged error estimates at a negligible computational overhead. Estimated standard deviations on the training data sets are compared to those from self-consistent calculations in Table IV. The ensemble performance on the data-set level should be assessed in combination with observing the error predictions on a system-to-system basis. Figure 6 illustrates the BEEF–vdW ensemble error estimates for the RE42 molecular reaction energies, and compares BEEF–vdW results to those of other functionals. Similar figures for more data sets are found in the Supplemental Material.[83]

On the data-set level, the overall predictive performance of the ensemble is satisfactory. The ensemble standard deviations in Table IV are slightly overestimated for the G2/97, CE17, and S22x5-0.9 data sets, while the ensemble underestimates the errors for RE42, DBH24/08, and Sol34Ec. For the remaining S22x5 subsets, the error estimates are too large.

Importantly, Fig. 6 illustrates strengths and weaknesses of the present approach to error estimation. Many of the reaction energies are accurately reproduced by BEEF–vdW, and the ensemble estimates a relatively small error on those data. However, some of the reactions for which BEEF–vdW yields larger errors are assigned too small error bars. The water-gas shift reaction $CO + H_2O \rightarrow CO_2 + H_2$ is one of these. The reason for this is indicated by the fact that all tested GGA, MGGA, and vdW–DF–type functionals yield nearly identical reaction energies for this reaction. One simply has to move rather far in XC model space to find a functional that predicts a reaction energy significantly different from the BEEF–vdW result. This causes the ensemble to underestimate the actual error for that reaction. Since the hybrid functionals appear to break the overall trends observed for the lower-rung functionals in Fig. 6, inclusion of exact exchange in the model space might remedy such limitations of the BEEF–vdW functional and its Bayesian ensemble.

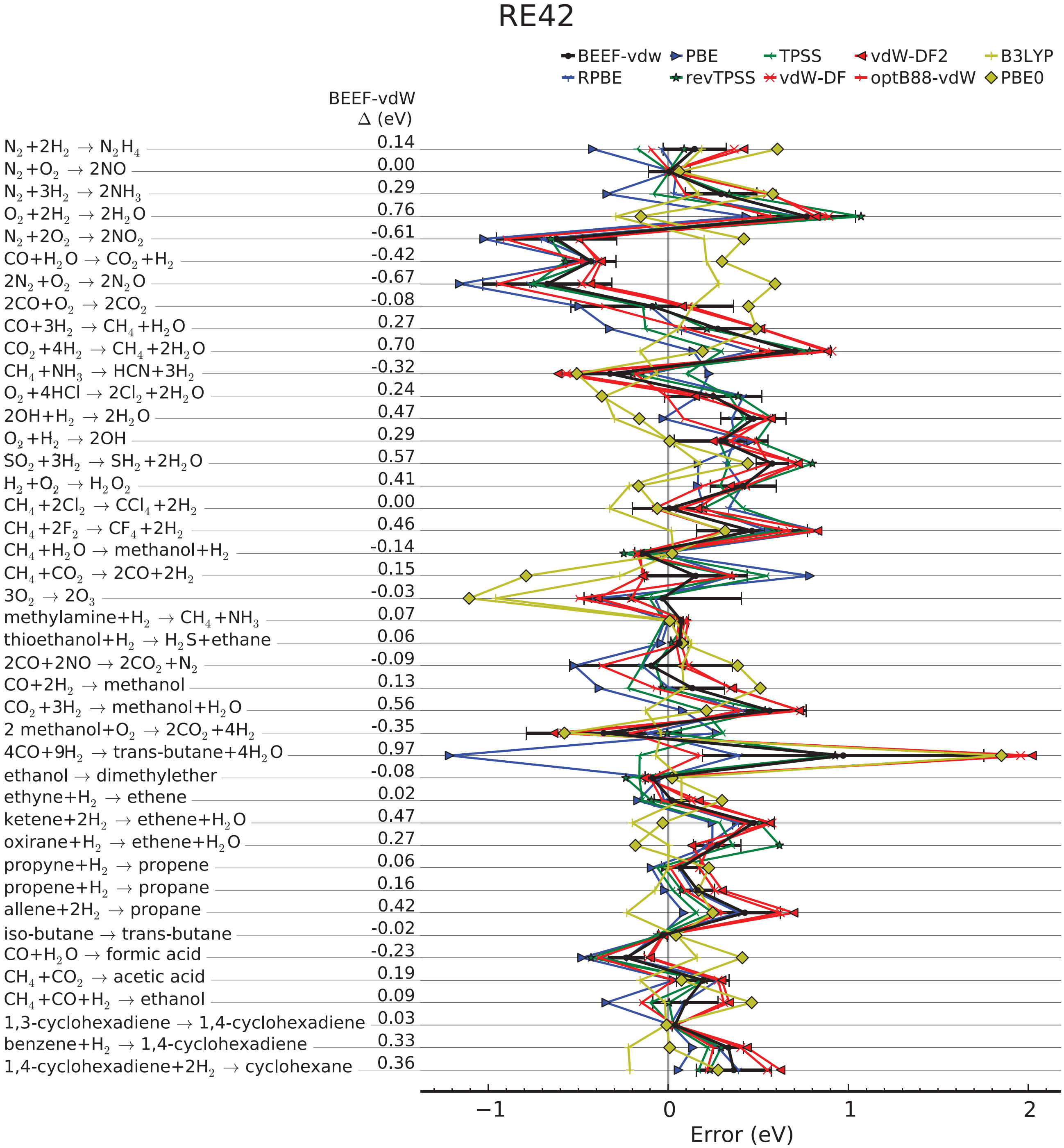


FIG. 6. (Color online) Deviations $\Delta = \Delta_r E^{\mathrm{DFT}} - \Delta_r E^{\mathrm{exp}}$ between the RE42 molecular reaction energies calculated using representative XC functionals and experiment. Color codes are BEEF–vdW: black; GGA: blue; MGGA: green; vdW–DF type: red; and hybrid: yellow. BEEF–vdW ensemble error estimates are indicated by horizontal error bars. The numbers in the middle column are self-consistent BEEF–vdW deviations from experiment.

VII. BENCHMARKS

The following is a comparative assessment of BEEF–vdW and a selection of literature XC functionals of the LDA, GGA, MGGA, vdW–DF, and hybrid types. These are listed in Table V. The benchmark data sets used are the six sets to which BEEF–vdW was trained, except Sol34Ec, as well as the G3-3, CE27, Sol27Ec, and Sol27LC data sets. The latter sets were introduced in Sec. II. Statistics on deviations of computed quantities from experimental or high-level theoretical references are reported for each density functional in terms of the mean signed (MSD), mean absolute (MAD), and standard deviation (STD). The sign

TABLE V. A selection of density functionals at the LDA (1), GGA (2), MGGA[a] (3), vdW–DF (3.5), and hybrid[b] (4) rungs of Jacob's ladder.

	Type	Target[c]	Ref.
LDA	1		68
PBE	2	General	67
RPBE	2	Chemistry	76
BLYP	2	Chemistry	87, 88
HCTH407	2	Chemistry	46
PBEsol	2	Solid state	43
WC	2	Solid state	89
AM05	2	Solid state	90
TPSS	3	General	91
revTPSS	3	General	44
vdW-DF	3.5	vdW	25
vdW-DF2	3.5	vdW	30
optPBE-vdW	3.5	vdW	33
optB88-vdW	3.5	vdW	33
C09-vdW	3.5	vdW	34
B3LYP	4	Chemistry	92
PBE0	4	Chemistry	93

[a] Attempts to apply the M06-L (Ref. 47) MGGA were unsuccessful due to convergence issues for a wide range of systems from almost all considered data sets. Note that problematics of evaluating MGGA potentials, especially for the M06 family of functionals, are discussed in recent literature (Refs. 84–86).

[b] Hybrid functionals have not been applied to extended systems.

[c] Should be understood as a very general characterization of the main target of a functional, and does not consider underlying principles of design.

convention is

$$\text{deviation} = \text{DFT} - \text{reference}. \tag{25}$$

Computed deviations for all systems in all data sets considered are tabulated in the Supplemental Material,[83] which also provides the raw DFT data.

All data are furthermore available online in the Computational Materials Repository (CMR).[95] The repository contains all information about the individual DFT calculations which form the basis for the results presented here, including atomic configurations and GPAW specific parameters. Access to search, browse, and download these data is provided through the CMR web interface.[96]

A. Molecular formation energies

The G2/97 and G3/99 thermochemical test sets have become standards for validating density functional methods, and the present calculations are well in line with published benchmark data[94] for these sets. Statistics are reported in Table VI. Considering first G2/97, the LDA grossly overestimates the molecular formation energies. Significant improvements are found with GGAs, where XC functionals designed to capture molecular energetics (RPBE, BLYP, HCTH407) yield STDs below 0.5 eV, while those targeted at solid state properties (PBEsol, WC, AM05) perform significantly worse: their MSDs are large and negative, indicating severe overbinding. The TPSS and revTPSS MGGA approximations perform quite well on this set.

Turning to the vdW–DF variants, good description of the G2/97 formation energies is also found for vdW–DF and vdW–DF2. This, however, is not the case for the optPBE–vdW, optB88–vdW, and C09–vdW functionals, for which the GGA exchange components are optimized with vdW dominated energetics in mind. This approach apparently leads to intramolecular overbinding, as previously noted in Ref. 31.

For comparison, Table VI also includes statistics for the B3LYP and PBE0 hybrids. As the wide application of hybrid XC functionals in the quantum chemistry community suggests, B3LYP and PBE0 accurately describe molecular bond energetics, and the B3LYP parametrization is found to be the best DFA for the G2/97 data set. Table VI furthermore shows that also the BEEF–vdW functional performs very well in predicting molecular formation energies. With a MAD of 0.16 eV, BEEF–vdW is highly accurate on the G2/97 thermochemical set, and even outperforms the PBE0 hybrid on these systems.

Now, let us switch attention to the G3-3 set of 75 molecules, which the BEEF–vdW model was not trained on. For most XC functionals tested here, the average deviations on G3-3 are larger than on G2/97. It is, however, noteworthy that TPSS, revTPSS, vdW–DF, and vdW–DF2 are exceptions to this trend. Benchmarking BEEF–vdW on G3-3 validates its good performance in predicting molecular bond energetics. This conclusion is underlined by the BEEF–vdW deviation statistics on the full G3/99 compilation. With a MAD of 0.19 eV, it is the most accurate DFA tested on G3/99, closely followed by B3LYP. Both MGGA functionals as well as vdW–DF and vdW–DF2 also perform well on this set.

B. Molecular reaction energies

The last column of Table VI summarizes deviation statistics for the RE42 data set. Even though the reaction energies are derived from the G2/97 formation energies, the reaction energies appear difficult to capture accurately with GGA, MGGA, and vdW–DF type functionals. None of them yield a STD less than 0.3 eV. The B3LYP hybrid proves significantly more accurate in this respect. Interestingly, the optPBE–vdW and optB88–vdW functionals, which both severely overestimate the G2/97 formation energies, prove as reliable for calculating gas-phase reaction energies as the best GGA (RPBE), and compare well to TPSS and BEEF–vdW.

C. Chemisorption on solid surfaces

Deviation statistics for the CE17 and CE27 data sets are reported in the first two columns of Table VII. The BEEF–vdW model was trained on CE17, while CE27 contains 10 extra entries, mostly covering dissociative H_2 chemisorption on late transition-metal surfaces. With MADs $\geqslant 0.7$ eV, LDA and the GGAs designed for solid state applications are clearly overbinding simple adsorbates to solid surfaces (negative MSDs). The RPBE, BLYP, and HCTH407 functionals are significantly more reliable for calculation of chemisorption energies, RPBE performing best with a MAD of 0.11 eV for both CE17 and CE27. Also, vdW–DF and vdW–DF2 yield

TABLE VI. Deviation statistics on the G2/97, G3-3, and G3/99 thermochemical data sets, as well as the RE42 set of molecular reaction energies. All energies in eV.

	G2/97 (148)			G3-3 (75)			G3/99 (223)			RE42 (42)		
Method	MSD	MAD	STD	MSD	MAD	STD	MSD	MAD	STD	MSD	MAD	STD
LDA	−3.69	3.69	4.27	−8.35	8.35	8.78	−5.25	5.25	6.16	−0.55	1.06	1.62
PBE	−0.64	0.68	0.84	−1.32	1.32	1.48	−0.87	0.90	1.10	−0.08	0.30	0.42
RPBE	0.25	0.40	0.51	0.94	0.96	1.13	0.48	0.59	0.78	0.11	0.26	0.34
PBEsol	−1.69	1.70	2.00	−3.94	3.94	4.14	−2.45	2.45	2.90	−0.29	0.48	0.73
BLYP	0.00	0.32	0.43	0.57	0.62	0.76	0.19	0.42	0.56	0.16	0.29	0.37
AM05	−1.77	1.78	2.07	−4.00	4.00	4.19	−2.52	2.52	2.96	−0.21	0.41	0.62
WC	−1.24	1.26	1.51	−2.86	2.86	3.03	−1.79	1.80	2.14	−0.24	0.43	0.65
HCTH407	0.09	0.26	0.35	0.48	0.55	0.65	0.22	0.36	0.47	0.06	0.27	0.35
TPSS	−0.22	0.28	0.33	−0.26	0.29	0.33	−0.24	0.28	0.33	0.06	0.25	0.32
revTPSS	−0.21	0.28	0.34	−0.24	0.26	0.31	−0.22	0.27	0.33	0.16	0.33	0.43
vdW–DF	−0.10	0.24	0.33	0.18	0.24	0.32	−0.01	0.24	0.33	0.24	0.39	0.52
vdW–DF2	−0.15	0.28	0.39	0.11	0.26	0.36	−0.06	0.28	0.38	0.24	0.40	0.54
optPBE–vdW	−0.84	0.85	0.98	−1.72	1.72	1.82	−1.14	1.14	1.32	0.06	0.27	0.35
optB88–vdW	−1.04	1.04	1.20	−2.22	2.22	2.34	−1.44	1.44	1.68	0.02	0.26	0.34
C09–vdW	−1.55	1.55	1.80	−3.55	3.55	3.72	−2.22	2.22	2.61	−0.11	0.33	0.45
B3LYP[a]	0.05	0.14	0.19	0.36	0.37	0.41	0.15	0.21	0.28	−0.05	0.15	0.22
PBE0[a]	−0.10	0.21	0.28	−0.40	0.44	0.55	−0.20	0.29	0.39	0.13	0.33	0.47
BEEF-vdW	−0.02	0.16	0.24	0.19	0.25	0.31	0.05	0.19	0.27	0.14	0.29	0.37

[a] B3LYP and PBE0 data adapted from Ref. 94.

MADs of 0.20 eV of less on CE27, while the two MGGAs overbind on average. Again, a significant overbinding is found for the three exchange-modified vdW–DF flavors. Lastly, it is seen from the CE17 column in Table VII that BEEF–vdW is among the DFAs offering most accurate predictions of chemisorption energies of simple adsorbates on solid surfaces. Since much of this accuracy is retained when moving to CE27, good transferability is expected when applying BEEF–vdW to other types of surface processes involving rupture and formation of chemical bonds.

D. Molecular reaction barriers

The DBH24/08 reaction barrier heights belong to a class of systems for which a fraction of exact exchange is known to increase computational accuracy significantly over GGAs.[22,97] This is supported by the DBH24/08 data in Table VII, where the two hybrids clearly outperform the lower-rung XC functionals. Considering the corresponding statistics for BEEF–vdW as well as for the individual DBH24/08 XC model reported in Table I, where a MAD of 0.12 eV was obtained, it is

TABLE VII. Deviation statistics on the CE17 and CE27 chemisorption energies, DBH24/08 reaction barriers, and the S22x5 interaction energies of noncovalently bonded complexes. All energies in eV, except S22x5, which is in meV.

	CE17 (17)			CE27 (27)			DBH24/08 (24)			S22x5 (110)		
Method	MSD	MAD	STD	MSD	MAD	STD	MSD	MAD	STD	MSD	MAD	STD
LDA	−1.34	1.34	1.39	−1.33	1.33	1.42	−0.58	0.58	0.73	−50	62	110
PBE	−0.42	0.42	0.44	−0.40	0.40	0.43	−0.33	0.33	0.43	76	76	132
RPBE	−0.02	0.11	0.13	0.00	0.11	0.14	−0.27	0.27	0.34	138	138	227
PBEsol	−0.85	0.85	0.87	−0.85	0.85	0.89	−0.44	0.44	0.56	38	53	85
BLYP	−0.04	0.13	0.16	0.02	0.15	0.18	−0.33	0.33	0.39	140	140	218
AM05	−0.70	0.70	0.73	−0.69	0.69	0.73	−0.41	0.41	0.53	99	99	157
WC	−0.76	0.76	0.78	−0.76	0.76	0.80	−0.41	0.41	0.52	56	63	105
HCTH407	0.11	0.17	0.22	0.15	0.20	0.30	−0.19	0.21	0.31	115	116	218
TPSS	−0.32	0.32	0.37	−0.34	0.34	0.41	−0.35	0.35	0.41	100	100	162
revTPSS	−0.38	0.38	0.43	−0.38	0.38	0.45	−0.35	0.35	0.41	92	92	141
vdW–DF	−0.05	0.12	0.14	0.04	0.18	0.22	−0.27	0.28	0.34	39	52	87
vdW–DF-2	−0.04	0.13	0.18	0.07	0.20	0.26	−0.30	0.31	0.37	31	33	61
optPBE–vdW	−0.39	0.39	0.42	−0.31	0.35	0.40	−0.33	0.33	0.41	−4	21	29
optB88–vdW	−0.52	0.52	0.56	−0.44	0.45	0.52	−0.37	0.37	0.45	3	10	15
C09–vdW	−0.78	0.78	0.81	−0.73	0.73	0.79	−0.41	0.41	0.50	−5	12	18
B3LYP							−0.17	0.17	0.21	111	111	180
PBE0							−0.13	0.15	0.19	71	71	124
BEEF–vdW	−0.08	0.12	0.14	−0.01	0.16	0.19	−0.26	0.26	0.33	42	50	88

clear that the BEEF–vdW model has moved significantly away from the part of model space favored by gas-phase reaction barrier heights. Nevertheless, BEEF–vdW is among the best nonhybrid functionals for such quantities.

E. Noncovalent interactions

The last column of Table VII lists deviation statistics for the S22x5 interaction energies. As previously found in several studies[30,33,53,59] of the original S22 data set, vdW dominated interactions are well described by vdW–DF type density functionals, especially those with an optimized exchange component. With MADs of 20 meV or less over all 110 points on the 22 potential-energy curves, the optPBE–vdW, optB88–vdW, and C09–vdW functionals prove highly accurate in this respect. The vdW–DF2 functional also captures vdW interactions well, but the positive MSD signifies that most of the deviations from the CCSD(T) reference energies stem from underbinding. For vdW–DF and BEEF–vdW, this is even more pronounced. None of the tested MGGA or hybrid DFAs convincingly capture vdW interactions. Only the most weakly gradient enhancing GGAs (PBEsol, WC, AM05) yield reasonable statistics. Taking into account the appreciable LDA overbinding of the S22x5 complexes, what appears to be GGA functionals capturing long-ranged dispersion is more likely a case of getting it right for the wrong reasons.

For completeness, Table VIII shows detailed S22x5 statistics for vdW–DF variants and BEEF–vdW. Although performing reasonably well on S22x5 as a whole, the vdW–DF, vdW–DF2, and BEEF–vdW functionals underestimate the intermolecular binding energies at shortened binding distances $d = 0.9$. Also, at $d = 1.0$ the exchange-modified vdW–DF flavors offer a better description, but the difference between the two groups is much reduced. Concerning computational accuracy, the vdW–DF2 MSD of 43 meV and MAD of 44 meV for S22x5-1.0 obtained here compare very well to the MSD and MAD of 40 and 41 meV, respectively, found in a recent study[59] for a revised S22 data set.

TABLE VIII. Detailed statistics on the deviations of calculated S22x5 interaction energies from CCSD(T) benchmarks using van der Waals density functionals in all five points along the intermolecular potential-energy curve. Mean signed and mean absolute deviations are in meV. Mean signed relative (MSRD) and mean absolute relative (MARD) deviations are also listed. Negatively signed deviation means overbinding on average.

Method	MSD	MAD	MSRD	MARD
$d = 0.9$				
vdW–DF	140	140	198%	198%
vdW–DF2	99	99	143%	143%
optPBE–vdW	29	31	28%	35%
optB88–vdW	17	19	26%	26%
C09–vdW	−13	21	−13%	35%
BEEF–vdW	136	137	214%	214%
$d = 1.0$				
vdW–DF	70	71	20%	25%
vdW–DF2	43	44	13%	15%
optPBE–vdW	−1	20	−9%	13%
optB88–vdW	5	13	3%	6%
C09–vdW	−3	13	1%	6%
BEEF–vdW	72	74	20%	28%
$d = 1.2$				
vdW–DF	4	32	−16%	23%
vdW–DF2	5	13	−2%	7%
optPBE–vdW	−25	28	−29%	30%
optB88–vdW	−4	13	−6%	9%
C09–vdW	−3	13	−8%	11%
BEEF–vdW	6	27	−12%	18%
$d = 1.5$				
vdW–DF	−13	15	−39%	40%
vdW–DF2	2	4	4%	6%
optPBE–vdW	−20	20	−44%	44%
optB88–vdW	−3	6	−12%	13%
C09–vdW	−6	11	−26%	28%
BEEF–vdW	−5	6	−13%	14%
$d = 2.0$				
vdW–DF	−4	4	−20%	20%
vdW–DF2	5	5	34%	34%
optPBE–vdW	−5	5	−20%	21%
optB88–vdW	1	2	3%	8%
C09–vdW	−2	2	−13%	15%
BEEF–vdW	2	3	27%	28%

F. Solid state properties

Table IX reports a summary of deviation statistics for calculations of lattice constants (Sol27LC) and cohesive energies (Sol27Ec). The lattice constant statistics are in clear favor of the PBEsol, AM05, WC, and revTPSS functionals. Their standard deviations are small and the MSDs are close to 0 Å. On average, however, these remarkably accurate predictions of equilibrium crystal volumes come at the price of overestimated cohesive energies.

The picture is opposite for vdW–DF and vdW–DF2. Lattice constants are overestimated and more so than with any other XC functional tested, vdW–DF2 yielding a standard deviation of 0.18 Å. Furthermore, those two DFAs notably underestimate cohesive energies. The less repulsive exchange functionals of the modified vdW–DF variants lead in general to statistics similar to those of PBE and TPSS for the two materials properties in question. These findings closely match those reported in recent studies[32,78,98–100] assessing the performance of GGA, MGGA, and vdW–DF type XC functionals for solid state properties.

Benchmarking finally BEEF–vdW, we find in Table IX that it performs reasonably well for cohesive energies and lattice constants, though still predicting softer crystal lattices than the optimized vdW–DF variants. With BEEF–vdW, these two bulk materials properties are, however, significantly closer to agreement with experiments than predictions by vdW–DF, vdW–DF2, and most of the GGAs designed mainly for chemistry.

VIII. APPLICATIONS

Two applications of BEEF–vdW to problems of current interest in the surface science community are here presented: graphene adsorption on the close-packed Ni(111) surface, and the trends observed when applying lower-rung density functionals in calculations of the binding energy of CO to

TABLE IX. Deviation statistics for the Sol27Ec cohesive energies (eV/atom) and Sol27LC lattice constants (Å). Zero-point vibrational effects have been removed from both experimental data sets.

	Sol27Ec (27)			Sol27LC (27)		
Method	MSD	MAD	STD	MSD	MAD	STD
LDA	0.89	0.89	1.08	−0.07	0.07	0.10
PBE	−0.10	0.27	0.38	0.05	0.06	0.07
RPBE	−0.54	0.58	0.71	0.11	0.11	0.13
PBEsol	0.43	0.45	0.63	−0.01	0.03	0.04
BLYP	−0.79	0.80	0.89	0.11	0.11	0.14
AM05	0.25	0.36	0.51	0.01	0.03	0.04
WC	0.37	0.41	0.57	0.00	0.03	0.04
HCTH407	−0.59	0.67	0.82	0.08	0.10	0.14
TPSS	0.08	0.27	0.36	0.05	0.05	0.08
revTPSS	0.31	0.37	0.50	0.03	0.04	0.07
vdW–DF	−0.54	0.60	0.72	0.12	0.12	0.14
vdW–DF2	−0.58	0.64	0.75	0.12	0.14	0.18
optPBE–vdW	−0.12	0.27	0.38	0.06	0.08	0.10
optB88–vdW	0.01	0.25	0.36	0.04	0.08	0.09
C09–vdW	0.42	0.43	0.59	0.01	0.05	0.06
BEEF–vdW	−0.37	0.45	0.59	0.08	0.08	0.11

Pt(111) and Rh(111) substrates as well as the surface energy of those substrates.

A. Graphene adsorption on Ni(111)

The remarkable electronic properties of monolayer graphene[103–105] and its potential application in electronics technology[104,106] motivate investigation of the interactions between graphene sheets and metallic surfaces. The nature of graphene adsorption on metals is highly metal dependent,[107,108] some surfaces binding graphene only weakly and others forming strong covalent bonds to the carbon sheet. The Ni(111) surface belongs to the latter group, graphene forming a (1 × 1) overlayer at a graphene-metal distance of $d = 2.1$ Å.[109] Furthermore, a band gap is induced in graphene upon adsorption, underlining the strong hybridization responsible for changing the electronic structure of the carbon sheet.[110,111]

Several theoretical studies have investigated the graphene/Ni(111) potential-energy curve, with mixed results.[112–118] However, based on RPA calculations, it is by now established that this particular adsorption process is a delicate competition between strong interactions close to the surface and vdW forces further from the surface.[101,102] Figure 7 shows calculated PECs for graphene adsorption on Ni(111) using LDA, MGGA, and vdW–DF type density functionals, as well as BEEF–vdW. Computational details are given in the Appendix. Additionally, two sets of RPA data are shown for comparison, indicating that graphene adsorption on Ni(111) is characterized by a physisorption minimum at $d = 3.0$–3.5 Å and a chemisorbed state at $d \approx 2.2$ Å, the latter in good agreement with experiments.[109] However, as previously found,[101,102,116,117] rung 1–3 DFAs, as well as vdW–DF and vdW–DF2, fail to simultaneously describe both qualitative features. Conversely, the optPBE–vdW and optB88–vdW PECs are increasingly closer to RPA data. The BEEF–vdW PEC shows qualitatively similar features, but the

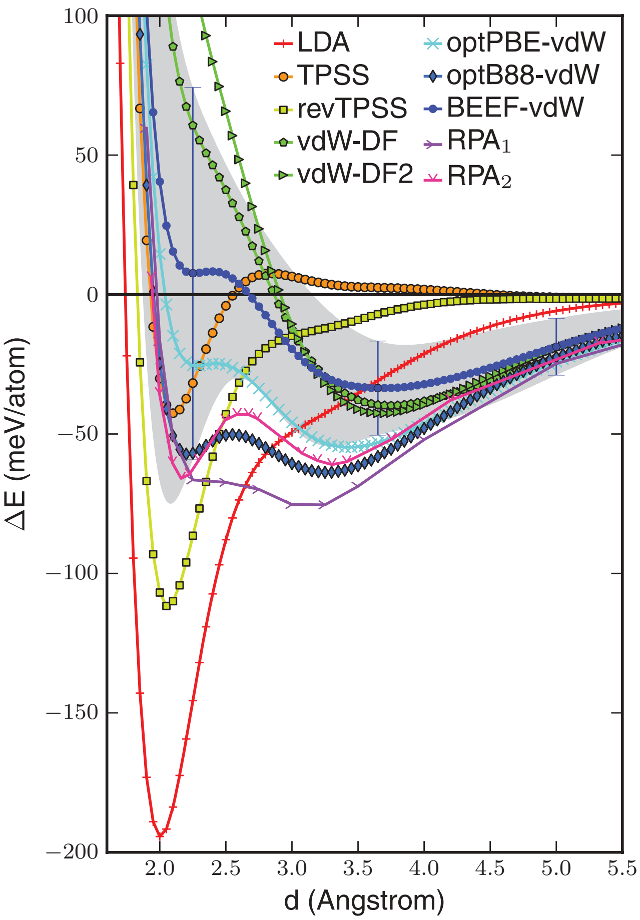


FIG. 7. (Color online) Potential-energy curves for graphene adsorption on the Ni(111) surface. Random phase approximation data are from Refs. 101 (RPA$_1$) and 102 (RPA$_2$). The gray area indicates the region spanned by the estimated standard deviations along the BEEF–vdW PEC.

local minimum at $d = 2.25$ Å is very shallow and yields a positive adsorption energy.

Figure 7 also shows ensemble error estimates along the BEEF–vdW PEC. Especially two aspects of these are of interest. First of all, the error bars do not straddle the zero line for large graphene-metal distances, indicating that confidence in the presence of a physisorption minimum is high. Second, the error bars enlarge notably at smaller distances from $d = 2.6$ Å and inwards, reflecting that these BEEF–vdW data points are associated with a significantly larger uncertainty. Recalling how the ensemble error estimate is designed (Sec. VI), the error estimates indicate that the graphene/Ni(111) PEC is very sensitive to the choice of XC functional in the chemically interesting range. Put differently, the ensemble suggests that we should not trust the BEEF–vdW prediction of a positive PEC for $d < 2.7$ Å as a definite result, as the estimated errors are simply too large in this region of the PEC.

B. Surface chemistry and stability

Chemisorption energies of molecules on surfaces are obviously important quantities in heterogeneous catalysis and surface science. However, accurate computation of surface energies E_γ can be critical as well since minimization of surface energy is a driving force determining the morphology and composition of surfaces, interfaces, and nanoparticles.[123] GGA density functionals, however, often underestimate E_γ, and the GGAs yielding most accurate surface energies also vastly overbind molecules to surfaces.[119] It thus appears that accurate computation of chemisorption energies on a surface as well as the stability of that surface is not possible with the same GGA approximation, underscoring a fundamental incompleteness of the GGA XC model space.

The issue is here investigated for vdW–DF variants and BEEF–vdW. Figure 8 shows atop chemisorption energies of CO on Pt(111) and Rh(111) against surface energies of those substrates, calculated using GGA, MGGA and vdW–DF type functionals, and BEEF–vdW with error estimation. These are compared to RPA results and experimental data. As previously reported,[119,124] the GGA data points fall along an approximately straight line, which is significantly offset from the experimental data, thus illustrating the issue discussed above. This is here shown to be the case for vdW–DF variants also: The dashed vdW–DF lines are parallel to the solid GGA lines, and are only slightly offset from the latter, especially for Rh(111). The vdW–DF and vdW–DF2 data points are quite close to RPBE. Larger surface energies are found with the exchange-modified vdW–DF variants, albeit at the expense of overestimated chemisorption energies. Note that such a correlation should be expected from Tables VII and IX and a linear relation between E_γ and the solid cohesive energy.[123]

Although BEEF–vdW contains the vdW–DF2 nonlocal correlation functional as an essential component, the former predicts larger surface energies than the latter without sacrificing accuracy of the CO-metal binding energy. We expect that this ability of BEEF–vdW to "break" the vdW–DF line is due to the expanded GGA model space as compared to vdW–DF, the latter of which pairs nonlocal correlation with LDA correlation. Significant inclusion of semilocal correlation in vdW–DF type calculations was also found in Ref. 31 to broadly improve accuracy for several materials properties. The BEEF–vdW error estimates furthermore appear very reasonable. The experimental CO chemisorption energies are straddled for both Pt(111) and Rh(111), and the error estimates along E_γ almost fill out the gap between the GGA lines to the left and the RPA and C09–vdW surface energies to the right. Lastly, it is seen from the green TPSS and revTPSS data points in Fig. 8, as also reported in Ref. 124, that the third rung of Jacob's ladder may offer the possibility of quite

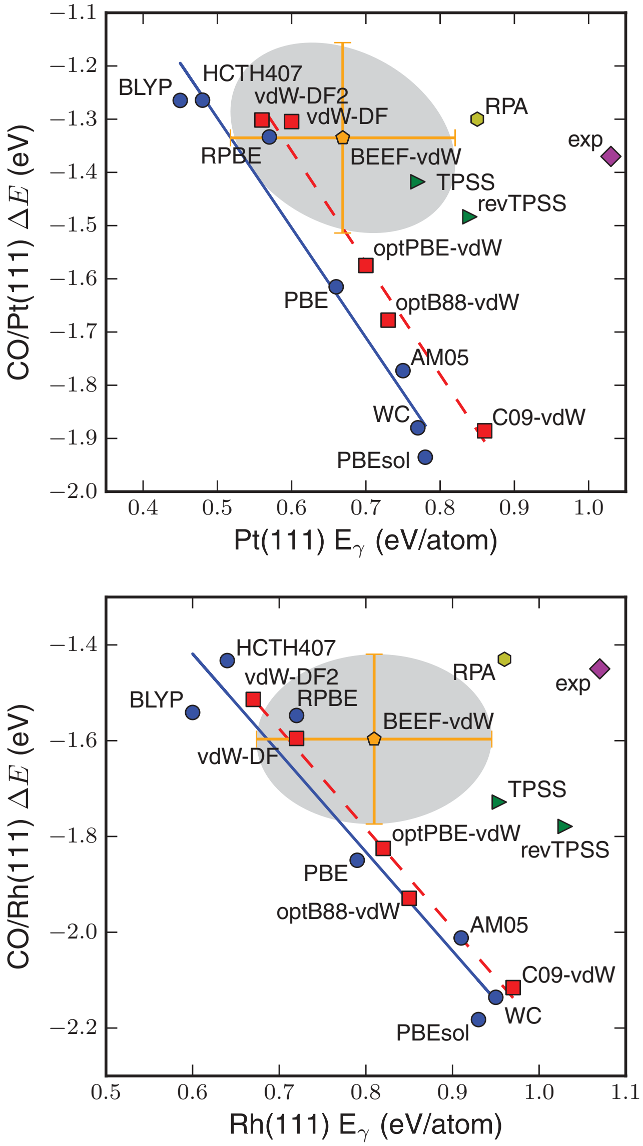


FIG. 8. (Color online) Atop CO chemisorption energies ΔE versus surface energies E_γ for Pt(111) and Ru(111). Red and blue lines are linear fits to GGA and vdW–DF type data points, respectively. MGGA data in green and yellow RPA data adapted from Ref. 119. Estimated standard deviations are indicated by error bars around the orange BEEF–vdW data points. All points $(E_\gamma, \Delta E)$ inside the gray areas are within one standard deviation from the BEEF–vdW point for both quantities. Experimental surface energies from liquid-metal data (Refs. 120 and 121), and experimental CO chemisorption energies from Ref. 122.

accurate surface energies with only moderately overbound surface adsorbates.

IX. DISCUSSION

The presented approach to semiempirical DFA development fundamentally considers XC functionals as more or less accurate models of the exact density functional. From this point of view, the XC model space expansion and model selection procedure are essential, as are data sets for calibrating or benchmarking XC models. The concept of an ensemble of model solutions is intrinsic to the present model selection procedure. The cost function for a single data set has both weak (sloppy) and strong (important) eigenmodes in a sufficiently flexible model space. Regularization is used to suppress the weak modes in order to facilitate a physically sensible model and maximize transferability. The regularized ensemble thus contracts around the strong modes, and the optimum model can, to some extent, be regarded an average of the ensemble solutions. Without Tikhonov regularization of exchange, all XC approximations obtained in this work would have 31 parameters and wildly oscillating GGA exchange solutions, corresponding to a least-squares fit of an order-30 polynomial in the reduced density gradient. Instead, well-behaved models with 3–8 effective parameters are obtained.

It is important to note that model selection is intricately connected to the model space. The reduced density gradient transformation $t(s)$ defines the expansion of GGA exchange. It thereby also determines how hard the regularization punishes nonsmoothness in different regions of s space, as well as how the exchange part of the prior solution transforms to s space. As previously stated, the prior is the origo for the XC model solution. Many different priors may be chosen, but we find it convenient that it transforms to a reasonable exchange approximation. Then, decreasing regularization from infinity towards zero leads to increasingly nonsmooth variations away from this initial guess.

The linear combination correlation model space of local, semilocal, and nonlocal correlation was anticipated[31] to enable highly accurate calculations for several, if not all, of the data sets considered. The individually trained models in Table I confirm this, some sets favoring full LDA correlation in addition to nonlocal ditto, other sets preferring full PBE correlation, while most sets are fitted best by a combination of both. The corresponding exchange functionals are also significantly different, so the sets of strong eigenmodes for the regularized cost functions are very materials property dependent. We argue here that explicitly considering transferability among different materials properties is important for producing a single DFA composed of the most important modes for the combined data sets, that is, the optimum model compromise must be found.

One approach to this task is minimizing a weighted sum of the individual cost functions. This is somewhat similar to weighted training functions used in least-squares-fitting procedures, but with the critically important addition of regularization. The summed cost function is elegantly minimized using the individual solutions only, but gives no information regarding how the weights should be chosen. Clearly, an XC model trade-off is inevitable, so the weights should be the ones yielding an optimum compromise. For just two data sets, a wide range of poor choices of weights can be made, and the complexity of this choice increases with the number of data sets. In line with the statistical approach taken in the bulk of this work, we believe that such choice should not be made based on experience or intuition alone. Rather, a systematic methodology for locating one or more points in XC model space, where a well-behaved and properly compromising solution resides, is desirable. The condition of minimizing the product of relative costs for each data set is a reasonable requirement for the model solution, philosophically as well as in practice: The condition essentially states that if changing the solution vector $\mathbf{a}$ to $\mathbf{a} + \delta\mathbf{a}$ gains a larger relative reduction in cost on one materials property than is lost in total on all other properties considered, then $\mathbf{a} + \delta\mathbf{a}$ is preferred.

Extensive benchmarking of BEEF–vdW against popular GGA, MGGA, vdW–DF type, and hybrid XC functionals shows that the developed methodology is able to produce truly general-purpose XC approximations. Results are summarized in Fig. 9, where error statistics for representative functionals on gas-phase chemical, surface chemical, solid state, and vdW dominated data sets are illustrated by bars. The BEEF–vdW model compromise is indeed a very agreeable one. For none of the data sets is the average BEEF–vdW error among the largest, while several other functionals are highly biased towards certain types of materials properties. This is especially true for vdW–DF2 and optB88–vdW, displaying severely erroneous description of binding energetics for bulk solids and molecules, respectively. Furthermore, the figure shows an overall performance equivalence of BEEF–vdW and the original vdW–DF for gas-phase and surface chemical properties, although the former more accurately predicts bonding in the solid state. Further testing of the functional might, however, prove interesting. Systems such as ionic solids, semiconductors, and transition-metal complexes are not included in the present benchmark, nor are the BEEF–vdW predictions of molecular ionization potentials and electron affinities tested. This will be addressed in future work.

We emphasize the strengths and weaknesses of the BEEF–vdW ensemble error estimate. The ensemble functionals are based on a probability distribution for the model parameters, which limits the ensemble to the BEEF–vdW model space only. This space is incomplete in the sense that it can not accommodate a physically reasonable XC model yielding zero error on all systems in all data sets considered, hence the model trade-off. The BEEF–vdW computational errors are in general reasonably well estimated, but the energetics of certain systems is rather insensitive to the choice of XC approximation within the GGA, MGGA, and vdW–DF type model spaces. This leads to relatively small error estimates for these systems, even though the actual computational error may be substantial.

Meanwhile, we find BEEF–vdW and the Bayesian ensemble highly useful in surface science related applications. The fact that BEEF–vdW appears to yield more accurate surface energies than GGA or vdW–DF type XC approximations of similar accuracy for adsorbate-surface bond strengths is very promising. The error estimate proves very useful in this case, even though the kinetic energy density of MGGA type functionals may be needed in the model space if the surface energy error bars are to span the experimental data. This again illustrates that the ensemble does not give information beyond

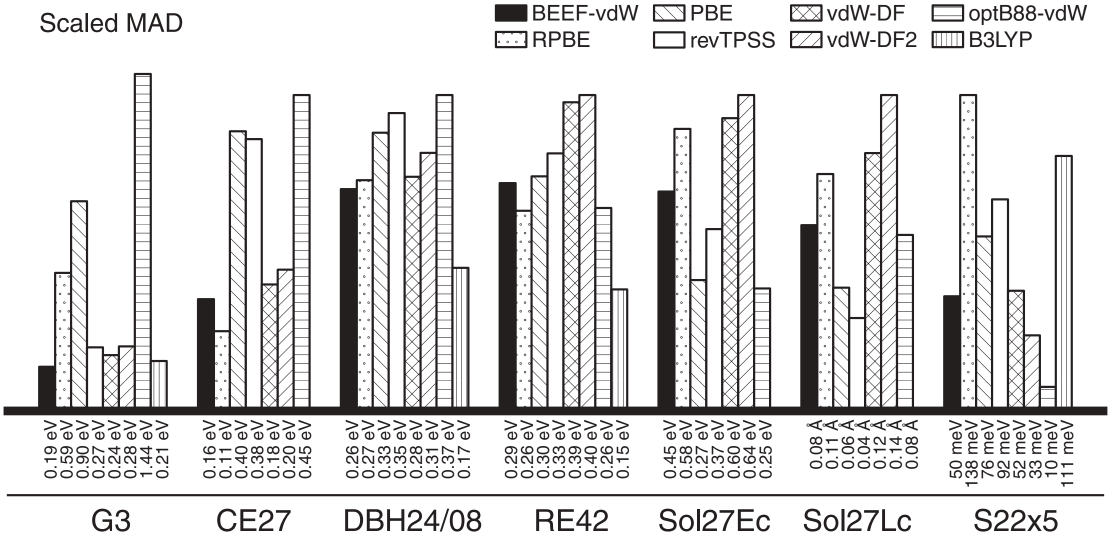


FIG. 9. Bar plot comparison of the accuracy of different density functionals in predicting various materials properties. For each data set, the bars illustrate proportionally scaled mean absolute deviations. The data sets are chosen to represent intramolecular bond energetics (G3), chemisorption energetics of molecules on surfaces (CE27), molecular reaction barrier heights (DBH24/08), molecular reaction energies (RE42), bulk solid cohesive energies (Sol27Ec) and lattice constants (Sol27LC), and interaction energies of noncovalently bonded complexes (S22x5). B3LYP calculations were not performed for bulk solids nor the extended CE27 systems.

its model space, as it is solely based on it. However, the error estimate carries important information in the BEEF–vdW study of graphene adsorption on Ni(111). The PEC is qualitatively wrong in the region of chemical bonding for this intricate case of "solid state adsorption," and the estimated errors indeed indicate that this part of the BEEF–vdW PEC is poorly determined. BEEF–vdW calculations can therefore not predict with any confidence whether graphene should form chemical bonds to the Ni(111) substrate in a low-temperature experiment. It is encouraging that the ensemble is able to capture this.

X. SUMMARY AND CONCLUSIONS

We have presented and evaluated a machine-learning-inspired approach to semiempirical density functional development. Focus has been on general applicability of the resulting density functional to both strong and weak interactions in chemistry and condensed matter physics, including surface chemistry. Transferability and avoiding overfitting are thus key issues, leading the presented methodology to rely primarily on (1) a variety of data sets chosen to represent vastly different interactions and bonding situations, (2) a very flexible XC model space expansion at a computationally feasible GGA + vdW level of approximation, and (3) XC model selection procedures designed to "tame" the flexible model space and yield XC approximations which properly compromise between describing different types of physics and chemistry.

To conclude, we have shown that regularization and cross-validation methods are very useful for semiempirical density functional development in highly flexible model spaces. It is furthermore clear that computationally efficient general-purpose functionals, targeted at accurately describing several physically and chemically different materials properties, necessarily must compromise between those properties in an incomplete XC model space. However, the optimum model trade-off is not easily found from simple intuition. A simple but powerful principle for determining the position in model space of a properly compromising XC approximation is therefore formulated.

Application of the developed methodology has yielded the BEEF–vdW density functional, and a benchmark of BEEF–vdW against popular GGA, MGGA, vdW–DF type, and hybrid XC functionals for energetics in chemistry and condensed matter physics has been conducted. This benchmark validates BEEF–vdW as a general-purpose XC approximation, with a reasonably reliable description of van der Waals forces and quantitatively accurate prediction of chemical adsorption energies of molecules on surfaces, while avoiding large sacrifices on solid state bond energetics. This should make it a valuable density functional for studies in surface science and catalysis.

Furthermore, an error estimation ensemble of functionals around BEEF–vdW comes out naturally of the developed fitting methodology. The ensemble is designed to provide an easily obtainable estimate of the XC approximation error. It is based on a probability distribution for the XC model parameters, and has been applied in the BEEF–vdW benchmark and qualitative assessments for molecular surface adsorption, surface energies, and graphene adsorption on Ni(111).

Finally, the methods developed here should lend themselves well to other XC model spaces also, including the MGGA level of theory or self-interaction correction schemes.

ACKNOWLEDGMENTS

The authors are grateful for the helpful suggestions and comments by N. Marzari, F. Abild-Pedersen, and J. VandeVondele. This work was supported by the Danish Center for

Scientific Computing. The Center for Atomic-Scale Materials Design is sponsored by the Lundbeck Foundation. The work at SUNCAT/SLAC has been supported by the US Department of Energy, Basic Energy Sciences.

APPENDIX: DETAILS OF DATA SETS AND COMPUTATIONS

1. G2/97 and G3/99

In accordance with the procedure of Ref. 49, the G3/99 formation enthalpies are corrected for thermal and vibrational contributions using thermal corrections and zero-point energies from Refs. 49 and 94. The G3/99 set is divided into three subsets denoted G3-1, G3-2, and G3-3 comprising 55, 93, and 75 molecules, respectively. The G3-1 and G3-2 subsets constitute G2/97. The G3-3 subset contains a significant fraction of larger carbon-rich molecules as compared to G2/97.

Theoretical G3/99 formation energies $\Delta_f E$ are calculated from the difference between molecular and atomic total energies as

$$\Delta_f E = E_M - \sum_A E_A, \tag{A1}$$

where A runs over all atoms in the molecule M, while E_M and E_A are ground-state molecular and atomic total energies at 0 K, respectively.

2. RE42

The 42 molecular reaction energies $\Delta_r E$ of the RE42 set are listed in Table X. Theoretical reaction energies are calculated from the total energies of G2/97 molecules after full geometry relaxation as

$$\Delta_r E = \sum_P E_P - \sum_R E_R, \tag{A2}$$

where the sums run over reactant (R) and product (P) molecules.

3. DBH24/08

Forward (V_f) and backward (V_b) benchmark reaction barriers from high-level theory or experiments are adapted from Ref. 50. Ground- and transition-state molecular geometries determined from quadratic configuration interaction calculations with single and double excitations (QCISD) are from Ref. 136. Density functional barrier heights are computed from the transition-state total electronic energy (E_{ts}) and the initial (E_i) and final (E_f) state total energies as

$$V_f = E_{ts} - E_i, \quad V_b = E_{ts} - E_f. \tag{A3}$$

4. S22x5

The original S22 publication[51] from 2006 reported CCSD(T) interaction energies of 22 noncovalently bonded complexes with extrapolation to the complete basis-set (CBS) limit. However, different basis sets were used for small and large complexes. Geometries were determined from MP2 or CCSD(T) calculations. Later works[61,137] have revised the S22 interaction energies, employing larger and identical basis sets for all complexes without changing the geometries. For the larger complexes, the reported basis-set effects are significant, so we use here the CCSD(T)/CBS energies of Takatani *et al.*[61] as the current best estimate of the true S22 interaction energies.

TABLE X. Gas-phase molecular reactions and reaction energies (in eV) constituting the RE42 data set. The experimental reaction energies are compiled from the G2/97 static-nuclei formation energies. $\Delta_r E < 0$ means exothermic.

Reaction	$\Delta_r E$
$N_2 + 2H_2 \rightarrow N_2H_4$	0.41
$N_2 + O_2 \rightarrow 2NO$	1.88
$N_2 + 3H_2 \rightarrow 2NH3$	−1.68
$O_2 + 2H_2 \rightarrow 2H_2O$	−5.45
$N_2 + 2O_2 \rightarrow 2NO_2$	0.62
$CO + H_2O \rightarrow CO_2 + H_2$	−0.31
$2N_2 + O_2 \rightarrow 2N_2O$	1.57
$2CO + O_2 \rightarrow 2CO_2$	−6.06
$CO + 3H_2 \rightarrow CH_4 + H_2O$	−2.80
$CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O$	−2.50
$CH_4 + NH_3 \rightarrow HCN + 3H_2$	3.32
$O_2 + 4HCl \rightarrow 2Cl_2 + 2H_2O$	−1.51
$2OH + H_2 \rightarrow 2H_2O$	−6.19
$O_2 + H_2 \rightarrow 2OH$	0.74
$SO_2 + 3H_2 \rightarrow SH_2 + 2H_2O$	−2.62
$H_2 + O_2 \rightarrow H_2O_2$	−1.68
$CH_4 + 2Cl_2 \rightarrow CCl_4 + 2H_2$	0.19
$CH_4 + 2F_2 \rightarrow CF_4 + 2H_2$	−8.60
$CH_4 + H_2O \rightarrow$ methanol $+ H_2$	1.33
$CH_4 + CO_2 \rightarrow 2CO + 2H_2$	3.11
$3O_2 \rightarrow 2O_3$	2.92
methylamine $+ H_2 \rightarrow CH_4 + NH_3$	−1.15
thioethanol $+ H_2 \rightarrow H_2S +$ ethane	−0.71
$2CO + 2NO \rightarrow 2CO_2 + N_2$	−7.94
$CO + 2H_2 \rightarrow$ methanol	−1.48
$CO_2 + 3H_2 \rightarrow$ methanol $+ H_2O$	−1.17
2 methanol $+ O_2 \rightarrow 2CO_2 + 4H_2$	−3.11
$4CO + 9H_2 \rightarrow$ trans-butane $+ 4H_2O$	−9.00
ethanol → dimethylether	0.53
ethyne $+ H_2 \rightarrow$ ethene	−2.10
ketene $+ 2H_2 \rightarrow$ ethene $+ H_2O$	−1.92
oxirane $+ H_2 \rightarrow$ ethene $+ H_2O$	−1.56
propyne $+ H_2 \rightarrow$ propene	−2.00
propene $+ H_2 \rightarrow$ propane	−1.58
allene $+ 2H_2 \rightarrow$ propane	−3.64
iso-butane → trans-butane	0.08
$CO + H_2O \rightarrow$ formic acid	−0.39
$CH_4 + CO_2 \rightarrow$ acetic acid	0.28
$CH_4 + CO + H_2 \rightarrow$ ethanol	−0.91
1,3-cyclohexadiene → 1,4-cyclohexadiene	−0.01
benzene $+ H_2 \rightarrow$ 1,4-cyclohexadiene	−0.01
1,4-cyclohexadiene $+ 2H_2 \rightarrow$ cyclohexane	−2.94

The S22x5 (Ref. 52) CCSD(T)/CBS potential-energy curves were reported more recently. The computational protocol was, however, not updated from that used for S22, so we expect the aforementioned interaction-energy inaccuracies to persist for S22x5. In order to shift the equilibrium point on each PEC to the revised S22 energies, and approximately correct the remaining data points, a modification of the (possibly) slightly inaccurate S22x5 CCSD(T) interaction energies is here introduced as

$$E^d_{\mathrm{int}} := \varepsilon^d_{\mathrm{int}} \times \frac{E^{1.0}_{\mathrm{int}}}{\varepsilon^{1.0}_{\mathrm{int}}}, \tag{A4}$$

where E^d_{int} and $\varepsilon^d_{\mathrm{int}}$ denote modified and original S22x5 energies at the relative intermolecular distance d, respectively. For $E^{1.0}_{\mathrm{int}} = \varepsilon^{1.0}_{\mathrm{int}}$, Eq. (A4) obviously reduces to $E^d_{\mathrm{int}} = \varepsilon^d_{\mathrm{int}}$ for all distances. The obtained corrections to $\varepsilon^d_{\mathrm{int}}$ are listed in Table XI. The maximum correction of 11.4% amounts to 25.6 meV for the indole-benzene complex in a stacked geometry, while the total mean signed correction to all the 110 interaction energies is 0.1 meV.

The modified CCSD(T) interaction energies are used throughout this study for the S22x5 data set and subsets. Each density functional interaction energy E^d_{int} is calculated as the difference between the total electronic energy of the interacting complex E^d_0 and those of its two isolated molecular constituents E^d_1 and E^d_2:

$$E^d_{\mathrm{int}} = E^d_0 - E^d_1 - E^d_2. \tag{A5}$$

Computational accuracy is enhanced by keeping all atoms in the molecular fragments in the same positions in the box as those atoms have when evaluating the total energy of the complex.

TABLE XI. Corrections $E^d_{\mathrm{int}} - \varepsilon^d_{\mathrm{int}}$ to the S22x5 interaction energies in Ref. 52 computed from Eq. (A4). Reported statistics are most negative (min), most positive (max), mean signed (msc), and mean absolute (mac) interaction energy correction at each distance. Furthermore, the total mean signed (MSC) and total mean absolute (MAC) energy corrections over all 110 energies are reported in the bottom rows. All energies in meV.

		Relative interaction distance d				
Complex	$E^{1.0}_{\mathrm{int}}/\varepsilon^{1.0}_{\mathrm{int}}$	0.9	1.0	1.2	1.5	2.0
1	−1.0%	−1.0	−1.3	−1.0	−0.5	−0.1
2	−1.0%	−1.9	−2.2	−1.8	−1.0	−0.4
3	−1.1%	−8.0	−9.1	−7.6	−4.5	−1.8
4	−1.1%	−6.5	−7.3	−6.1	−3.7	−1.6
5	−1.1%	−9.2	−10.0	−8.4	−5.1	−2.2
6	−1.8%	−11.8	−13.0	−10.8	−6.4	−2.5
7	−2.3%	−14.7	−16.0	−13.0	−7.3	−2.5
8	0.0%	0.0	0.0	0.0	0.0	0.0
9	−1.2%	−0.4	−0.8	−0.4	−0.1	0.0
10	3.2%	1.5	2.1	1.6	0.7	0.2
11	6.8%	0.4	8.3	5.7	1.6	0.2
12	6.9%	5.1	13.5	9.0	2.9	0.6
13	1.3%	3.8	5.6	3.6	1.4	0.4
14	11.4%	10.5	25.6	17.8	5.3	0.5
15	4.6%	15.9	24.3	16.4	6.5	1.8
16	−1.4%	−0.7	−0.9	−0.7	−0.3	−0.1
17	−0.6%	−0.8	−0.9	−0.7	−0.4	−0.1
18	1.3%	1.1	1.3	1.0	0.5	0.2
19	−0.7%	−1.2	−1.3	−1.1	−0.6	−0.2
20	3.2%	3.1	3.9	3.1	1.6	0.5
21	2.1%	4.5	5.2	4.4	2.5	1.0
22	−0.6%	−1.6	−1.8	−1.5	−0.9	−0.4
min	−2.3%	−14.7	−16.0	−13.0	−7.3	−2.5
max	11.4%	15.9	25.6	17.8	6.5	1.8
msc	1.2%	−0.5	1.1	0.4	−0.4	−0.3
mac	2.5%	4.7	7.0	5.3	2.4	0.8
MSC	0.1					
MAC	4.0					

TABLE XII. Experimental solid-state properties of 27 cubic bulk solids. The ZPAE exclusive Sol27LC 0-K lattice constants a_0 (Å) are adapted from Ref. 98. 0-K Sol27Ec cohesive energies E_c (eV/atom) from Ref. 125 are corrected for ZPVE contributions. Strukturbericht symbols are indicated in parentheses for each solid. A1: fcc, A2: bcc, A3: hcp, A4: diamond.

	Sol27LC	Sol27Ec	
Solid	a_0	E_c	ZPVE[a]
Li (A2)	3.451	1.66	0.033
Na (A2)	4.209	1.13	0.015
K (A2)	5.212	0.94	0.009
Rb (A2)	5.577	0.86	0.005
Ca (A1)	5.556	1.86	0.022
Sr (A1)	6.040	1.73	0.014
Ba (A2)	5.002	1.91	0.011
V (A2)	3.024	5.35	0.037
Nb (A2)	3.294	7.60	0.027
Ta (A2)	3.299	8.12	0.023
Mo (A2)	3.141	6.86	0.044
W (A2)	3.160	8.94	0.039
Fe (A2)	2.853	4.33	0.046
Rh (A1)	3.793	5.80	0.047
Ir (A1)	3.831	6.98	0.041
Ni (A1)	3.508	4.48	0.044
Pd (A1)	3.876	3.92	0.027
Pt (A1)	3.913	5.86	0.023
Cu (A1)	3.596	3.52	0.033
Ag (A1)	4.062	2.97	0.022
Au (A1)	4.062	3.83	0.016
Pb (A1)	4.912	2.04	0.010
Al (A1)	4.019	3.43	0.041
C (A4)	3.544	7.59	0.216
Si (A4)	5.415	4.69	0.063
Ge (A4)	5.639	3.89	0.036
Sn (A4)	6.474	3.16	0.019

[a] ZPVE corrections are calculated according to Eq. (A6) using Debye temperatures from Ref. 125.

5. Sol27LC and Sol27Ec

It was recently shown[78] that removal of thermal and zero-point contributions to experimentally determined lattice constants and bulk moduli may be important when benchmarking density functional methods. Experimental zero Kelvin lattice constants and cohesive energies (E_c) contain zero-point vibrational contributions, leading to zero-point anharmonic expansion (ZPAE) of the lattice and zero-point vibrational energy (ZPVE) contributions to E_c. As discussed in Ref. 138, an estimate of the ZPVE may be obtained from the Debye temperature Θ_D of the solid according to

$$\mathrm{ZPVE} = -\tfrac{9}{8} k_B \Theta_D. \tag{A6}$$

The vibrational contribution is subtracted from the cohesive energy, leading to increased stability of the crystal towards atomization. The same reference derived a semiempirical estimate of the ZPAE contribution to the volume of cubic crystals. A recent study[18] calculating the ZPAE from first principles largely validates this approach. The Sol27LC and Sol27Ec sets of zero Kelvin lattice constants and cohesive energies of 27 fcc, bcc, and diamond structured bulk solids are appropriately corrected for zero-point phonon effects. Details are given in Table XII.

Density functional computation of total energies of the extended bulk solids is done using a 16×16×16 **k**-point mesh for sampling reciprocal space of the periodic lattice and 0.1 eV Fermi smearing of the electron occupation numbers. Calculations for bulk Fe, Ni, and Co are spin polarized. The cohesive energy for a given crystal lattice constant a is calculated from

$$E_c = E_A - E_B, \tag{A7}$$

where E_A is the total energy of the free atom and E_B is the bulk total energy per atom. By this definition, the equilibrium cohesive energy of a stable solid is a positive quantity. Equilibrium lattice constants of cubic crystals a_0 are determined from fitting the stabilized jellium equation of state (SJEOS, Ref. 138) to cohesive energies sampled in five points in a small interval around the maximum of the $E_c(a)$ curve.

6. CE17 and CE27

The CE17 and CE27 data are derived from temperature-programed desorption experiments or from microcalorimetry, most often at low coverage. The 27 chemisorption energies have been critically chosen from literature with emphasis on reliability as well as covering a reasonably wide range of substrates and adsorbates. All data are listed in Table XIII along with details regarding adsorption mode, adsorption site, and references.

Most of the CE27 surface reactions are molecular adsorption processes at 0.25 ML coverage. In that case, the chemisorption energy is computed according to

$$\Delta E = E_{AM} - E_M - xE_A, \tag{A8}$$

where E_{AM} is the total electronic energy of the adsorbate A on metal surface M, and E_A and E_M total energies of the isolated adsorbate and metal surface, respectively. The constant x equals 1 for molecular adsorption and N_2 dissociation on Fe(100), while $x = \frac{1}{2}$ for the dissociative H_2 chemisorption reactions. In the case of NO dissociation on Ni(100) at 0.25 ML coverage, the chemisorption energy is

$$\Delta E = E_{AM} + E_{BM} - 2E_M - E_{AB}, \tag{A9}$$

where AB is the NO molecule.

With these definitions of chemisorption energies, we consider extended surface slab models with 2×2 atoms in each layer and five layers in total. The slab models are periodic in the surface plane and a vacuum width of 20 Å separates periodically repeated slabs perpendicularly to the surface planes. Calculations involving Fe, Ni, and Co are spin polarized. Well-converged chemisorption energies are obtained using a 10 × 10 × 1 **k**-point mesh and a real-space grid spacing around 0.16 Å. The self-consistently determined lattice constant of the slab solid obviously determines the xy dimensions of the slab simulation cell. Since the number of real-space grid points employed in each direction is discrete, a grid spacing of exactly 0.16 Å in the x and y directions is rarely possible for slab calculations. Instead, it may be slightly smaller or larger, which should not affect the computational accuracy significantly. During structure relaxations, the bottom two layers of the 2 × 2 × 5 slab models are fixed in the bulk structure as found from bulk calculations.

TABLE XIII. The 27 experimental reaction energies ΔE for chemisorption on late transition-metal surfaces constituting the CE27 data set. The somewhat smaller CE17 data set is a subset of CE27. Reactions in CE17 are marked with a "⋆". All chemisorption energies are in eV per adsorbate at a surface coverage of 0.25 ML, except where otherwise noted. The adsorption mode is indicated by "m" (molecular) or "d" (dissociative), along with the adsorption site. Chemisorption energies for O have been evaluated as $\frac{1}{2}\{\Delta E(O_2) - E_b(O_2)\}$ with $E_b(O_2) = 118$ kcal/mol (Ref. 126) for the dioxygen bond energy.

		Mode	Site	ΔE	Reference(s)
CO/Ni(111)	⋆	m	fcc	−1.28	122
CO/Ni(100)		m	hollow	−1.26	127
CO/Rh(111)	⋆	m	top	−1.45	122
CO/Pd(111)	⋆	m	fcc	−1.48	122
CO/Pd(100)	⋆	m	bridge	−1.60	127–130
CO/Pt(111)	⋆	m	top	−1.37	122
CO/Ir(111)	⋆	m	top	−1.58	122
CO/Cu(111)	⋆	m	top	−0.50	122
CO/Co(0001)	⋆	m	top	−1.20	122
CO/Ru(0001)	⋆	m	top	−1.49	122
O/Ni(111)	⋆	m	fcc	−4.95	130
O/Ni(100)	⋆	m	hollow	−5.23	130
O/Rh(100)	⋆	m	hollow	−4.41	130
O/Pt(111)	⋆	m	fcc	−3.67	131
NO/Ni(100)	⋆	d	hollow	−3.99	127
NO/Pd(111)	⋆	m	fcc	−1.86	132
NO/Pd(100)	⋆	m	hollow	−1.61	133
NO/Pt(111)		m	fcc	−1.45	131
N_2/Fe(100)[b]		d	hollow	−2.3	134
H_2/Pt(111)	⋆	d	fcc	−0.41	135
H_2/Ni(111)		d	fcc	−0.98	135
H_2/Ni(100)		d	hollow	−0.93	135
H_2/Rh(111)		d	fcc	−0.81	135
H_2/Pd(111)		d	fcc	−0.91	135
H_2/Ir(111)		d	fcc	−0.55	135
H_2/Co(0001)		d	fcc	−0.69	135
H_2/Ru(0001)[c]		d	fcc	−1.04	135

[a] ΔE is the average of −1.58, −1.67, −1.69, and −1.45 eV.

[b] The coverage of atomic nitrogen is 0.5 ML.

[c] ΔE is the average of −0.83 and −1.24 eV, both from Ref. 135.

7. Graphene adsorption on Ni(111)

Adsorption of graphene on Ni(111) was modeled using a 1 × 1 × 5 surface slab, a Ni(fcc) lattice constant of 3.524 Å as determined with the PBE density functional, and 20 Å vacuum width. The top three atomic layers were fully relaxed with PBE using a grid spacing of 0.15 Å and a (20 × 20 × 1) **k**-point mesh. Carbon atoms were placed in atop and fcc adsorption sites, respectively.

*jesswe@fysik.dtu.dk

[1] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).

[2] W. Kohn, A. D. Becke, and R. G. Parr, J. Phys. Chem. **100**, 12974 (1996).

[3] W. Kohn, Rev. Mod. Phys. **71**, 1253 (1999).

[4] E. A. Carter, Science **321**, 800 (2008).

[5] B. Hammer and J. K. Nørskov, in *Impact of Surface Science on Catalysis*, Advances in Catalysis, Vol. 45, edited by B. C. Gates and H. Knözinger (Academic, New York, 2000), p. 71.

[6] M. J. Field, J. Comput. Chem. **23**, 48 (2002).

[7] J. Greeley, J. K. Nørskov, and M. Mavrikakis, Annu. Rev. Phys. Chem. **53**, 319 (2002).

[8] J. K. Nørskov, T. Bligaard, A. Logadottir, S. Bahn, L. B. Hansen, M. V. Bollinger, H. S. Bengaard, B. Hammer, Z. Sljivancanin, M. Mavrikakis, Y. Xu, S. Dahl, and C. J. H. Jacobsen, J. Catal. **209**, 275 (2002).

[9] H. Grönbeck, Top. Catal. **28**, 59 (2004).

[10] K. M. Neyman and F. Illas, Catal. Today **105**, 2 (2005).

[11] R. A. Friesner and V. Guallar, Annu. Rev. Phys. Chem. **56**, 389 (2005).

[12] J. K. Nørskov, M. Scheffler, and H. Toulhoat, MRS Bull. **31**, 669 (2006).

[13] H. M. Senn and W. Thiel, Angew. Chem. Int. Ed. **48**, 1198 (2009).

[14] M. Neurock, J. Catal. **216**, 73 (2003).

[15] J. Greeley and M. Mavrikakis, Nat. Mater. **3**, 810 (2004). .

[16] J. K. Nørskov, T. Bligaard, J. Rossmeisl, and C. H. Christensen, Nat. Chem. **1**, 37 (2009).

[17] J. K. Nørskov, F. Abild-Pedersen, F. Studt, and T. Bligaard, Proc. Natl. Acad. Sci. USA **108**, 937 (2011).

[18] L. Schimka, J. Harl, and G. Kresse, J. Chem. Phys. **134**, 024116 (2011).

[19] C. D. Sherrill, J. Chem. Phys. **132**, 110902 (2010).

[20] O. A. Vydrov, G. E. Scuseria, and J. P. Perdew, J. Chem. Phys. **126**, 154109 (2007).

[21] I. Dabo, A. Ferretti, N. Poilvert, Y. Li, N. Marzari, and M. Cococcioni, Phys. Rev. B **82**, 115121 (2010).

[22] A. J. Cohen, P. Mori-Sánchez, and W. Yang, Chem. Rev. **112**, 289 (2012).

[23] J. P. Perdew and K. Schmidt, in *Density Functional Theory and its Application to Materials*, AIP Conference Proceedings, Vol. 577, edited by V. Van Doren, C. Van Alsenoy, and P. Geerlings (AIP, New York, 2001), p. 1.

[24] J. P. Perdew, A. Ruzsinszky, J. M. Tao, V. N. Staroverov, G. E. Scuseria, and G. I. Csonka, J. Chem. Phys. **123**, 062201 (2005).

[25] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist, Phys. Rev. Lett. **92**, 246401 (2004).

[26] A. D. Becke and E. R. Johnson, J. Chem. Phys. **124**, 014104 (2006).

[27] A. Tkatchenko and M. Scheffler, Phys. Rev. Lett. **102**, 073005 (2009).

[28] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, J. Chem. Phys. **132**, 154104 (2010).

[29] D. C. Langreth, B. I. Lundqvist, S. D. Chakarova-Kack, V. R. Cooper, M. Dion, P. Hyldgaard, A. Kelkkanen, J. Kleis, L. Kong, S. Li, P. G. Moses, E. Murray, A. Puzder, H. Rydberg, E. Schröder, and T. Thonhauser, J. Phys.: Condens. Matter **21**, 084203 (2009).

[30] K. Lee, E. D. Murray, L. Kong, B. I. Lundqvist, and D. C. Langreth, Phys. Rev. B **82**, 081101 (2010).

[31] J. Wellendorff and T. Bligaard, Top. Catal. **54**, 1143 (2011).

[32] J. Klimes, D. R. Bowler, and A. Michaelides, Phys. Rev. B **83**, 195131 (2011).

[33] J. Klimes, D. R. Bowler, and A. Michaelides, J. Phys.: Condens. Matter **22**, 022201 (2010).

[34] V. R. Cooper, Phys. Rev. B **81**, 161104(R) (2010).

[35] E. D. Murray, D. C. Lee, and K. Langreth, J. Chem. Theory Comput. **5**, 2754 (2009).

[36] F. O. Kannemann and A. D. Becke, J. Chem. Theory Comput. **5**, 719 (2009).

[37] B. G. Janesko, J. Chem. Phys. **133**, 104103 (2010).

[38] B. G. Janesko and A. Aguero, J. Chem. Phys. **136**, 024111 (2012).

[39] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).

[40] F. A. Hamprecht, A. J. Cohen, D. J. Tozer, and N. C. Handy, J. Chem. Phys. **109**, 6264 (1998).

[41] H. L. Schmider and A. D. Becke, J. Chem. Phys. **108**, 9624 (1998).

[42] T. Van Voorhis and G. E. Scuseria, J. Chem. Phys. **109**, 400 (1998).

[43] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, Phys. Rev. Lett. **100**, 136406 (2008).

[44] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, L. A. Constantin, and J. Sun, Phys. Rev. Lett. **103**, 026403 (2009).

[45] A. D. Becke, J. Chem. Phys. **107**, 8554 (1997).

[46] A. D. Boese and N. C. Handy, J. Chem. Phys. **114**, 5497 (2001).

[47] Y. Zhao and D. G. Truhlar, J. Chem. Phys. **125**, 194101 (2006).

[48] L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople, J. Chem. Phys. **112**, 7374 (2000).

[49] L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople, J. Chem. Phys. **106**, 1063 (1997).

[50] J. Zheng, Y. Zhao, and D. G. Truhlar, J. Chem. Theory Comput. **5**, 808 (2009).

[51] P. Jurecka, J. Sponer, J. Cerny, and P. Hobza, Phys. Chem. Chem. Phys. **8**, 1985 (2006).

[52] L. Grafova, M. Pitonak, J. Rezac, and P. Hobza, J. Chem. Theory Comput. **6**, 2365 (2010).

[53] A. Gulans, M. J. Puska, and R. M. Nieminen, Phys. Rev. B **79**, 201105(R) (2009).

[54] F. O. Kannemann and A. D. Becke, J. Chem. Theory Comput. **6**, 1081 (2010).

[55] L. A. Burns, A. Vazquez-Mayagoitia, B. G. Sumpter, and C. D. Sherrill, J. Chem. Phys. **134**, 084107 (2011).

[56] F. Hanke, J. Comput. Chem. **32**, 1424 (2011).

[57] L. Goerigk and S. Grimme, Phys. Chem. Chem. Phys. **13**, 6670 (2011).

[58] Y. Zhao and D. G. Truhlar, Theor. Chem. Account. **120**, 215 (2008).

[59] O. A. Vydrov and T. Van Voorhis, J. Chem. Phys. **133**, 244103 (2010).

[60] S. Grimme, S. Antony, S. Ehrlich, and H. Krieg, J. Chem. Phys. **132**, 154104 (2010).

[61] T. Takatani, E. G. Hohenstein, M. Malagoli, M. S. Marshall, and C. D. Sherrill, J. Chem. Phys. **132**, 144104 (2010).

[62] J. J. Mortensen, L. B. Hansen, and K. W. Jacobsen, Phys. Rev. B **71**, 035109 (2005).

[63] J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dulak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. A. Hansen, H. H. Kristoffersen, M. Kuisma, A. H. Larsen, L. Lehtovaara, M. Ljungberg, O. Lopez-Acevedo, P. G. Moses, J. Ojanen, T. Olsen, V. Petzold, N. A. Romero, J. Stausholm-Møller, M. Strange, G. A. Tritsaris, M. Vanin, M. Walter, B. Hammer, H. Häkkinen,

G. K. H. Madsen, R. M. Nieminen, J. K. Nørskov, M. Puska, T. T. Rantala, J. Schiøtz, K. S. Thygesen, and K. W. Jacobsen, J. Phys.: Condens. Matter **22**, 253202 (2010).

[64] The GPAW and ASE codes are available as parts of the CAMPOS software: [http://www.camd.dtu.dk/Software].

[65] P. E. Blöchl, Phys. Rev. B **50**, 17953 (1994).

[66] S. R. Bahn and K. W. Jacobsen, Comput. Sci. Eng. **4**, 56 (2002).

[67] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[68] J. P. Perdew and Y. Wang, Phys. Rev. B **45**, 13244 (1992).

[69] G. Román-Pérez and J. M. Soler, Phys. Rev. Lett. **103**, 096102 (2009).

[70] J. Wellendorff, A. Kelkkanen, J. J. Mortensen, B. I. Lundqvist, and T. Bligaard, Top. Catal. **53**, 378 (2010).

[71] E. H. Lieb and S. Oxford, Int. J. Quantum Chem. **19**, 427 (1981).

[72] J. P. Perdew, in *Electronic Structure of Solids '91*, Vol. 17, edited by P. Ziesche and H. Eschrig (Akademie Verlag, Berlin, 1991), p. 11.

[73] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. (Springer, New York, 2006).

[74] K. S. Brown and J. P. Sethna, Phys. Rev. E **68**, 021904 (2003).

[75] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, New York, 2009).

[76] B. Hammer, L. B. Hansen, and J. K. Nørskov, Phys. Rev. B **59**, 7413 (1999).

[77] V. Petzold, T. Bligaard, and K. W. Jacobsen, Top. Catal. (to be published).

[78] G. I. Csonka, J. P. Perdew, A. Ruzsinszky, P. H. T. Philipsen, S. Lebegue, J. Paier, O. A. Vydrov, and J. G. Angyan, Phys. Rev. B **79**, 155107 (2009).

[79] B. Efron, J. Am. Stat. Assoc. **78**, 316 (1983).

[80] If one or more of the individual cost functions have very strong modes, several stationary points of the product cost function can exist. In this case, one must carefully determine all stationary points, and select the one which represents the best compromise. For the subsequent fitting of the BEEF–vdW, this was a minor issue.

[81] J. J. Mortensen, K. Kaasbjerg, S. L. Frederiksen, J. K. Nørskov, J. P. Sethna, and K. W. Jacobsen, Phys. Rev. Lett. **95**, 216401 (2005).

[82] D. Sivia and J. Skilling, *Data Analysis: A Bayesian Tutorial* (Oxford University Press, Oxford, UK, 2006).

[83] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevB.85.235149 for (1) figures comparing the performance of representative XC functionals at system level for all data sets considered in benchmarking, and including BEEF–vdW ensemble error estimates, (2) tables of those quantities, which are also used for the benchmark statistics, (3) comma-separated values (CSV) files containing the raw DFT data for benchmarking, and (4) the BEEF–vdW $\mathbf{\Omega}^{-1}$ ensemble matrix in CSV format. All the CSV files contain header information.

[84] E. R. Johnson, R. A. Wolkow, and G. A. DiLabio, Chem. Phys. Lett. **394**, 334 (2004).

[85] E. R. Johnson, A. D. Becke, C. D. Sherrill, and G. A. DiLabio, J. Chem. Phys. **131**, 034111 (2009).

[86] S. E. Wheeler and K. N. Houk, J. Chem. Theory Comput. **6**, 395 (2010).

[87] A. D. Becke, Phys. Rev. A **38**, 3098 (1988).

[88] C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988).

[89] Z. Wu and R. E. Cohen, Phys. Rev. B **73**, 235116 (2006).

[90] R. Armiento and A. E. Mattsson, Phys. Rev. B **72**, 085108 (2005).

[91] J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria, Phys. Rev. Lett. **91**, 146401 (2003).

[92] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, J. Phys. Chem. **98**, 11623 (1994).

[93] C. Adamo and V. Barone, J. Chem. Phys. **110**, 6158 (1999).

[94] V. N. Staroverov, G. E. Scuseria, J. M. Tao, and J. P. Perdew, J. Chem. Phys. **119**, 12129 (2003).

[95] D. D. Landis, J. S. Hummelshøj, S. Nestorov, J. Greeley, M. Dułak, T. Bligaard, J. K. Nørskov, and K. Jacobsen, Comput. Sci. Eng. **99** (2012), doi:10.1109/MCSE.2012.16.

[96] The CMR web-interface is found at [http://cmr.fysik.dtu.dk].

[97] A. D. Boese and J. M. L. Martin, J. Chem. Phys. **121**, 3405 (2004).

[98] P. Haas, F. Tran, and P. Blaha, Phys. Rev. B **79**, 085104 (2009).

[99] J. Sun, M. Marsman, G. I. Csonka, A. Ruzsinszky, P. Hao, Y.-S. Kim, G. Kresse, and J. P. Perdew, Phys. Rev. B **84**, 035117 (2011).

[100] P. Hao, Y. Fang, J. Sun, G. I. Csonka, P. H. T. Philipsen, and J. P. Perdew, Phys. Rev. B **85**, 014111 (2012).

[101] T. Olsen, J. Yan, J. J. Mortensen, and K. S. Thygesen, Phys. Rev. Lett. **107**, 156401 (2011).

[102] F. Mittendorfer, A. Garhofer, J. Redinger, J. Klimes, J. Harl, and G. Kresse, Phys. Rev. B **84**, 201401(R) (2011).

[103] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, Science **306**, 666 (2004).

[104] A. K. Geim and K. S. Novoselov, Nat. Mater. **6**, 183 (2007).

[105] A. H. Castro Neto, F. Guinea, N. M. R. Peres, K. S. Novoselov, and A. K. Geim, Rev. Mod. Phys. **81**, 109 (2009).

[106] F. Schwierz, Nat. Nanotechnol. **5**, 487 (2010).

[107] J. Wintterlin and M.-L. Bocquet, Surf. Sci. **603**, 1841 (2009).

[108] Y. S. Dedkov and M. Fonin, New J. Phys. **12**, 125004 (2010).

[109] Y. Gamo, A. Nagashima, M. Wakabayashi, M. Teria, and C. Oshima, Surf. Sci. **374**, 61 (1997).

[110] A. Grüneis and D. V. Vyalikh, Phys. Rev. B **77**, 193401 (2008).

[111] A. Varykhalov, J. Sánchez-Barriga, A. M. Shikin, C. Biswas, E. Vescovo, A. Rybkin, D. Marchenko, and O. Rader, Phys. Rev. Lett. **101**, 157601 (2008).

[112] G. Bertoni, L. Calmels, A. Altibelli, and V. Serin, Phys. Rev. B **71**, 075402 (2005).

[113] G. Kalibaeva, R. Vuilleumier, S. Meloni, A. Alavi, G. Ciccotti, and R. Rosei, J. Phys. Chem. B **110**, 3638 (2006).

[114] G. Giovannetti, P. A. Khomyakov, G. Brocks, V. M. Karpan, J. van den Brink, and P. J. Kelly, Phys. Rev. Lett. **101**, 026803 (2008).

[115] M. Fuentes-Cabrera, M. I. Baskes, A. V. Melechko, and M. L. Simpson, Phys. Rev. B **77**, 035405 (2008).

[116] M. Vanin, J. J. Mortensen, A. K. Kelkkanen, J. M. Garcia-Lastra, K. S. Thygesen, and K. W. Jacobsen, Phys. Rev. B **81**, 081408 (2010).

[117] I. Hamada and M. Otani, Phys. Rev. B **82**, 153412 (2010).

[118] C. Gong, G. Lee, B. Shan, E. M. Vogel, R. M. Wallace, and K. Cho, J. Appl. Phys. **108**, 123711 (2010).

[119] L. Schimka, J. Harl, A. Stroppa, A. Grüneis, M. Marsman, F. Mittendorfer, and G. Kresse, Nat. Mater. **9**, 741 (2010).

[120] W. R. Tyson and W. A. Miller, Surf. Sci. **62**, 267 (1977).

[121] L. Vitos, A. V. Ruban, H. L. Skriver, and J. Kollar, Surf. Sci. **411**, 186 (1998).

[122] F. Abild-Pedersen and M. P. Andersson, Surf. Sci. **601**, 1747 (2007).

[123] I. Chorkendorff and J. W. Niemantsverdriet, *Concepts of Modern Catalysis and Kinetics*, 2nd ed. (Wiley-VCH, Weinheim, 2007).

[124] J. Sun, M. Marsman, A. Ruzsinszky, G. Kresse, and J. P. Perdew, Phys. Rev. B **83**, 121410(R) (2011).

[125] C. Kittel, *Indtroduction to Solid State Physics*, 8th ed. (Wiley, New York, 2005).

[126] H. B. Gray, *Chemical Bonds: An Introduction to Atomic and Molecular Structure* (University Science Books, Mill Valley, California, 1994).

[127] W. A. Brown, R. Kose, and D. A. King, Chem. Rev. **98**, 797 (1998).

[128] H. Conrad, G. Ertl, J. Koch, and E. E. Latta, Surf. Sci. **43**, 462 (1974).

[129] R. J. Behm, K. Christmann, and G. Ertl, J. Chem. Phys. **73**, 2984 (1980).

[130] Q. Ge, R. Kose, and D. A. King, Adv. Catal. **45**, 207 (2000).

[131] V. Fiorin, D. Borthwick, and D. A. King, Surf. Sci. **603**, 1360 (2009).

[132] R. D. Ramsier, Q. Gao, H. N. Neergaard Waltenburg, K.-W. Lee, O. W. Nooij, L. Lefferts, and J. T. Yates Jr., Surf. Sci. **320**, 209 (1994).

[133] Y. Y. Yeo, L. Vattuone, and D. A. King, J. Chem. Phys. **106**, 1990 (1997).

[134] F. Bozso, G. Ertl, M. Grunze, and M. Weiss, J. Catal. **49**, 18 (1977).

[135] K. Christmann, Surf. Sci. Rep. **9**, 1 (1988).

[136] J. Zheng, Y. Zhao, and D. G. Truhlar, J. Chem. Theory Comput. **3**, 569 (2007).

[137] R. Podeszwa, K. Patkowski, and K. Szalewicz, Phys. Chem. Chem. Phys. **12**, 5974 (2010).

[138] A. B. Alchagirov, J. P. Perdew, J. C. Boettger, R. C. Albers, and C. Fiolhais, Phys. Rev. B **63**, 224115 (2001).

Paper II

mBEEF: An accurate semi-local Bayesian error estimation density functional

Jess Wellendorff, Keld T. Lundgaard, Karsten W. Jacobsen, and Thomas Bligaard

Submitted to The Journal of Chemical Physics

mBEEF: An accurate semi-local Bayesian error estimation density functional

Jess Wellendorff,[1, 2, a] Keld T. Lundgaard,[1, 3] Karsten W. Jacobsen,[3] and Thomas Bligaard[1, 2]
[1] *SUNCAT Center for Interface Science and Catalysis, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA*
[2] *Department of Chemical Engineering, Stanford University, Stanford, CA 94305, USA*
[3] *Center for Atomic-scale Materials Design (CAMD), Department of Physics, Building 307, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark*

(Dated: 29 January 2014)

We present a general-purpose meta-generalized gradient approximation (MGGA) exchange-correlation functional generated within the Bayesian error estimation functional framework [Wellendorff *et al.*, Phys. Rev. B **85**, 235149 (2012)]. The functional is designed to give reasonably accurate density functional theory (DFT) predictions of a broad range of properties in materials physics and chemistry, while exhibiting a high degree of transferability. Particularly, it improves upon solid cohesive energies and lattice constants over the BEEF-vdW functional without compromising high performance on adsorption and reaction energies. We thus expect it to be particularly well-suited for studies in surface science and catalysis. An ensemble of functionals for error estimation in DFT is an intrinsic feature of exchange-correlation models designed this way, and we show how the Bayesian ensemble may provide a systematic analysis of the reliability of DFT based simulations.

I. INTRODUCTION

Electronic structure theory offers key insights into the properties of materials, chemical reactions, and biomolecules. Kohn-Sham density functional theory[1,2] (KS-DFT) has proven a powerful framework for electronic structure studies,[3] particularly due to a favorable tradeoff between the computational speed and accuracy that can be obtained within this theory. Density functional methods have over the past decade reached a level of maturity where they can be applied not just in detailed theoretical studies of a given material, but be used to search for novel materials for technologically relevant applications in materials science[4–10] and chemical engineering.[11–16] Such studies often take a screening approach where massive amounts of DFT calculations are performed using efficient semi-local approximations for the KS exchange-correlation (XC) energy and potential. These include generalized gradient approximations (GGAs) and recent meta-GGA (MGGA) functionals.[17,18]

The reliability of semi-local density functional approximations (DFAs) is, however, unfortunately not universal. No such single functional appears to offer sovereign accuracy with zero bias in prediction of materials properties across the board of condensed matter and chemistry.[19–22] The GGA and MGGA exchange-correlation model spaces are flexible but incomplete and can not accommodate an approximation that represents the exact XC functional in all aspects of practical importance. The result is an exchange-correlation model compromise on accuracy between different chemical and materials properties. However, semi-local DFT remains a favorite workhorse method within many research areas, so useful XC model compromises are warranted. Semi-empirical optimization lends itself well as a method for finding reasonably accurate compromises, but will never completely eliminate DFT errors. The BEEF class of functionals generalizes the fitting procedure for XC functionals to allow for estimation of the errors on the quantities calculated from density functional theory. The traditional assumption underlying functional fitting is that a "best-fit" exchange-correlation model fitted to a suitable set of systems might be transferable, meaning that it hopefully calculates the properties well for systems not included in the training data. The generalization of this concept, which underlies BEEF-type functionals, is, that if one defines an "optimal" ensemble of exchange-correlation models, such that the ensemble on average reproduces errors on the training data, then the errors predicted by a well-constructed ensemble could be transferable. The ensemble can then be used to estimate computational uncertainties on calculations for systems not included in the fitted data set.

We have in Ref. 23 established a semi-empirical framework for developing model-compromise optimized density functionals with error ensembles as a practical implementation of the ideas proposed in Refs. 24 and 25. That study led to the first practically useful Bayesian error estimation functional, the BEEF-vdW, containing a somewhat expensive non-local correlation term. This BEEF framework uses machine learning tools to find the optimal compromise between model complexity and model accuracy for a fitted general-purpose DFA in a highly flexible exchange-correlation model space. It furthermore uses ideas from Bayesian statistics[24] to construct an ensemble of XC functionals directly from the cost function that was minimized to find the optimally accurate and transferable exchange-correlation functional. This subsequently allows for fast and systematic error estimates on simulated quantities, as the ensemble is applied non-selfconsistently on the electron density that results from utilizing the optimally fitted functional. A number of surface science and catalysis studies[26–29] have successfully applied the BEEF-vdW functional, and have demonstrated significant improvements over traditionally

[a] Electronic mail: jewe@slac.stanford.edu

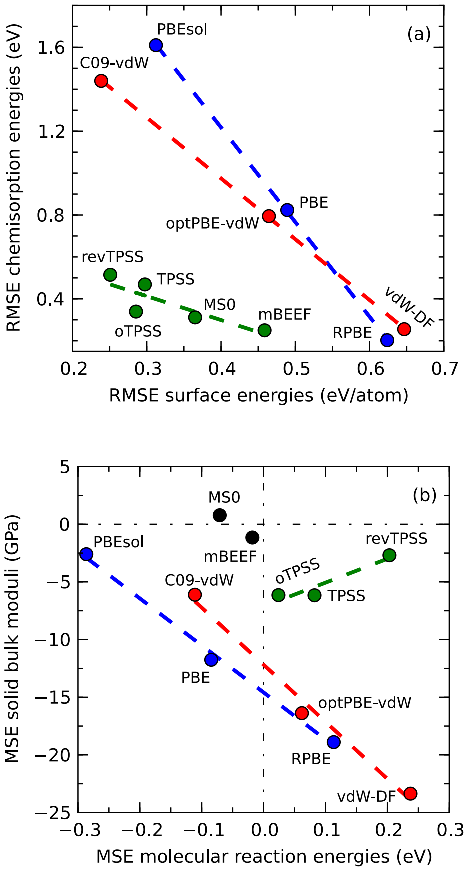


FIG. 1. Bivariate analyses of DFT prediction errors on chemical and materials properties. a) Root-mean-squared errors on the CE27a chemisorption energies against those on the SE30 surface energies. b) Mean-signed errors on the BM32 solid bulk moduli against those on the RE42 molecular reaction energies. Straight lines are fits through the GGA (blue), meta-GGA (green), and vdW-DF type (red) data. The meta-GGA lines are closest to origo in both panels, indicating improved possibilities for exchange-correlation model compromises in the meta-GGA model space as compared to the GGA and vdW-DF ones.

used GGAs[30] for similar studies.

We here take the BEEF development an important step forward by considering a meta-GGA exchange model space and refine the approach to XC model selection. This semi-local functional allows studies of larger and more complex systems than the BEEF-vdW, since the non-local correlation term has been eliminated. We shall show that the endured loss of accuracy, even for van der Waals-bonded systems, is rather limited. This work thus establishes the currently most versatile error estimation functional, particularly useful for systems that are not dominated by long-range dispersion interactions. We first illustrate the model compromise of semi-local DFT. A refined BEEF model selection procedure for addressing the model compromise is then introduced and applied to MGGA exchange. and the resulting density functional (mBEEF) is subsequently benchmarked. Finally, we illustrate the BEE approach to error estimation for a materials property in DFT by analyzing the adsorption-site preference of CO adsorption on late transition metal surfaces.

II. EXCHANGE-CORRELATION MODEL COMPROMISE

The Perdew-Burke-Ernzerhof (PBE)[31] approximation from 1996 has become a default GGA in many branches of the computational materials science research area. However, the hundreds of GGA functionals reported in literature since then clearly indicate that no GGA can be considered truly universal. The PBEsol[32] modification of PBE for example predicts bulk lattice constants remarkably well but severely overestimates molecular bond energies, while the RPBE[30] revision of the PBE functional describes covalent chemistry well at the expense of overestimated lattice constants and underestimated cohesive energies.[23,32,33] This is the topography of the XC model compromise in GGA DFT.

Meta-GGA density functionals[34–40] augment the GGA model space of electron density and its first-order gradient by including also the second-order density gradient[34] or the orbital kinetic energy density (KED) of the occupied KS eigenstates.[35] Importantly, an electronic structure with vanishing electron density gradient is in MGGA not necessarily modeled as a uniform electron gas (UEG). The UEG limit on exchange for small density gradients does in general not appear to be compatible with semi-local DFAs fully optimized for prediction of molecular bond energies.[23,25,41] Special-purpose GGAs may be designed by modification of known GGA forms, as in the cases of PBEsol and RPBE. The main purpose of applying MGGA exchange in the BEEF framework is, however, the prospect of better XC model compromises than with GGAs at a very modest increase in computational cost.[38]

We illustrate this point in Fig. 1, where a broad selection of GGAs, MGGAs, and vdW-DF[42] type functionals are applied in calculations of four different quantities; chemisorption energies of small molecules on close-packed transition metal facets, surface energies of various facets, solid bulk moduli, and gas-phase reaction energies. These properties are represented by the CE27a, SE30, BM32, and RE42 data sets, respectively, all discussed in more detail later. The tested GGAs are PBEsol, PBE, and RPBE, while the literature MGGAs are TPSS,[37] revTPSS,[38] oTPSS,[43] and MS0.[40] Note that the three representative van der Waals functionals vdW-DF, optPBE-vdW,[44] and C09-vdW[45] are equivalent except for the choice of PBE-like exchange. Figure 1a shows root-mean-squared errors (RMSEs) on the CE27a chemisorption energies against those on SE30 surface energies. The points within each class of XC model space fall approximately on straight lines, illustrating the trade-off one is forced to make between accurate adsorbate–surface bond strengths and surface stabilities. However, the MGGA model space offers the most attractive compromises; the green line in Fig. 1a is significantly closer to the origo. This is supported by Fig. 1b, in which mean-signed errors (MSEs) on predicted BM32 bulk moduli are plotted against those on RE42 molecular reaction energies. The relations between mean errors are again approximately linear and the MGGA points fall closest to origo, though not all on the same straight line.

The bivariate prediction error analysis in Fig. 1 confirms the conjectures from earlier studies[46,47] that the exchange-correlation model compromise of typical XC models lead directly to a trade-off between the systematic errors on various chemical and materials properties. Simple but efficient approaches to optimizing density functionals with respect to this trade-off are therefore core components of BEEF-class functional developments.

III. EXCHANGE MODEL SPACE

The spin-unpolarized meta-GGA exchange energy we write as the usual[37] integral over the uniform electron gas exchange energy density $\epsilon_x^{\mathrm{UEG}}$ scaled with a semi-local MGGA exchange enhancement factor F_x,

$$E_x = \int n\epsilon_x^{\mathrm{UEG}}(n)\; F_x(n,\nabla n,\tau)\; d\boldsymbol{r}, \tag{1}$$

where $n = n(\boldsymbol{r})$ is the local electron density, ∇n the density gradient, and the semi-local kinetic energy density $\tau = \frac{1}{2}\sum_{i,\sigma}|\nabla\Psi_{i,\sigma}|^2$ is summed over spins σ and state labels i of the KS eigenstates $\Psi_{i,\sigma}$. Atomic units are used throughout. The exchange enhancement factor we furthermore express in terms of dimensionless electronic structure parameters; the reduced density gradient $s = |\nabla n|/(2k_{\mathrm{F}}n)$, where $k_{\mathrm{F}} = (3\pi^2 n)^{\frac{1}{3}}$, and the reduced kinetic energy density $\alpha = (\tau-\tau^{\mathrm{W}})/\tau^{\mathrm{UEG}}$, where $\tau^{\mathrm{W}} = |\nabla n|^2/8n$ and $\tau^{\mathrm{UEG}} = (3/10)(3\pi^2)^{\frac{2}{3}}n^{\frac{5}{3}}$.

The MGGA exchange enhancement factor we therefore write $F_x(n,\nabla n,\tau) = F_x(s,\alpha)$, and expand it in products P of Legendre polynomials B depending on s and α through transformed quantities t_s and t_α:

$$t_s(s) = \frac{2s^2}{q+s^2} - 1, \tag{2}$$

$$t_\alpha(\alpha) = \frac{(1-\alpha^2)^3}{1+\alpha^3+\alpha^6}, \tag{3}$$

$$P_{mn} = B_m(t_s)B_n(t_\alpha), \tag{4}$$

$$F_x(s,\alpha) = \sum_{m=0}^{M}\sum_{n=0}^{N} a_{mn}P_{mn}. \tag{5}$$

For the mBEEF fit we chose values of $M = N = 7$, giving $Z = (M+1)\times(N+1) = 64$ exchange basis functions with expansion coefficients a_{mn}, which more than exhaust the present exchange model space. Both t_s and t_α are confined to $[-1,+1]$. With $q = \kappa/\mu = 0.804/(10/81) = 6.5124$, transformation t_s is a Padé approximant to the PBEsol $F_x(s)$, while t_α is inspired by the MS0 exchange.[40]

Denoting by E_x^{mn} the exchange energy corresponding to P_{mn}, the full exchange-correlation energy is written

$$\begin{aligned} E_{xc} &= \sum_{m,n}^{M,N} a_{mn}E_x^{mn} + E_c^{\mathrm{PBEsol}}, \\ &= \boldsymbol{x}\boldsymbol{a}^T + E_c^{\mathrm{PBEsol}}, \end{aligned} \tag{6}$$

where $\boldsymbol{x}$ is the vector of exchange basis function energy contributions for the system in question and the vector $\boldsymbol{a}$ contains the exchange model expansion coefficients in Eq. (5). The training data in $\boldsymbol{x}$ was obtained from PBEsol ground-state electron densities and single-particle eigenstates.

IV. TRAINING DATA SETS

Five significantly different sets of target chemical and materials properties are used in exchange model training. They cover in total a large part of the electronic structure phase space of molecules and condensed matter. Most of the sets were also applied in Ref. 23, but are here updated or slightly modified.

The G3/99 molecular formation energies[48] and the related RE42 reaction energies[23] represent gas-phase chemistry. Both data sets are normalized in model training such as to approximately bring all data within each set on an equal footing, see Ref. 49 for details. Surface chemistry we represent by the CE27a chemisorption energies of simple adsorbates on late transition metal surfaces adapted from previous work.[23,50] Solid bulk energetics is represented by cohesive energies in the Sol54Ec set, and bulk structures by the derivatives of cohesive energies with respect to crystal volume around equilibrium. Note that solid Pb is excluded from both data sets in model training, see Ref. 50. Experimental lattice constants are from the Sol58LC set.[50]

Density functional calculations are performed using GPAW,[51,52] an open-source DFT code implementing the projector augmented-wave method,[53] and the open-source ASE[54] package. GPAW can represent the Kohn-Sham equations on a real-space uniform grid as well as in a plane-wave expansion. Structural relaxations follow the prescriptions in Ref. 23 and use grid-point spacings of 0.15–0.16 Å. Chemisorption energies are calculated using a $(10\times10\times1)$ Monkhorst-Pack[55] $\boldsymbol{k}$-point mesh. Bulk calculations are done in plane-wave mode using a 1000 eV plane-wave energy cutoff and a $(16\times16\times16)$ $\boldsymbol{k}$-point mesh. Lattice constants and bulk moduli are computed by fitting the SJEOS equation of state[56] to 9 electronic total energies sampled at lattice constants spanning ±1% around the apparent equilibrium one.

V. EXCHANGE MODEL SELECTION

We seek a general-purpose density functional for surface chemistry studies with built-in error estimates. With the flexible exchange model space defined in Eq. (5), maximizing not only performance on the training data sets but also transferability to unseen data is essential. To this end we use ideas from machine learning[57,58] and extend on developments in Refs. 23 and 25. We formulate the optimization problem in terms of a regularized cost function to be minimized for the optimum (mBEEF) exchange coefficient vector $\hat{\boldsymbol{a}}_0$. We also generate a Bayesian error estimation ensemble in terms of model fluctuations around $\hat{\boldsymbol{a}}_0$. Several aspects of model selection are most conveniently introduced in terms of fitting a single set of data.

1. Cost function and BEE ensemble

We parametrize an exchange model on a single training set by minimizing a cost function C consisting of a squared-error loss function L and a regularizer R,

$$\begin{aligned} C(\boldsymbol{a};\omega) &= L(\boldsymbol{a}) + R(\boldsymbol{a};\omega) \\ &= \boldsymbol{q}^T\boldsymbol{q} + \omega^2\boldsymbol{b}^T\boldsymbol{b}, \end{aligned} \tag{7}$$

which depends parametrically on the regularization strength $\omega \geq 0$. The residual vector of training errors is $\boldsymbol{q} = \boldsymbol{X}\boldsymbol{a} - \boldsymbol{y}$, where matrix $\boldsymbol{X}$ contains all exchange basis function contributions and $\boldsymbol{y}$ is a vector of targets. The vector $\boldsymbol{b}$ is a suitable affine mapping of $\boldsymbol{a}$, which we shall define later in Eq. (15). The minimal cost solution $\boldsymbol{a}_0$ for a given choice of ω is easily found, see Ref. 23. In the language of Bayesian statistics, minimizing C over $\boldsymbol{a}$ given ω is equivalent to maximizing the posterior probability for the model parameters given a prior expectation.[57,58] The regularizer in Eq. (7) imposes the prior expectation for $\boldsymbol{a}_0$ as a penalty term of variable strength. The effect is parameter shrinkage, a standard machine learning method for dealing with ill-posed regression problems and avoiding over-fitting by controlling the model complexity.[57,58] Any ordinary least-squares (OLS) regression solution in a sufficiently large model space will contain a number of poorly determined parameters—parameters that vary wildly for small perturbations of the training data—a clear indication of over-fitting. Singular value decomposition of regularized cost functions of the form (7) shows how the regularizer adds curvature to such weak modes in L and essentially freezes them out of the fit.[25,58] Regularization is therefore used to tune the model complexity in order to enhance model generalization. It is then natural to introduce the notion of an effective number of model parameters θ,[57,58]

$$\theta(\omega) = \sum_m \frac{\nu_m}{\nu_m + \omega^2}, \tag{8}$$

where ν_m are the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$. Note that $\theta(0) = Z$ recovers the OLS solution while $\theta(\infty) = 0$. We may think of θ as the number of cost function eigenmodes that are not significantly affected by regularization.[25,57]

The cost function is quadratic in $\boldsymbol{a}$ and can therefore around its minimum $C_0 = C(\boldsymbol{a}_0;\omega_0)$ be expressed in terms of the Hessian matrix $\boldsymbol{H} = \partial^2 C/\partial\boldsymbol{a}^T\partial\boldsymbol{a}$ and model perturbations $\delta\boldsymbol{a} = \boldsymbol{a} - \boldsymbol{a}_0$:

$$C(\boldsymbol{a}) = C_0 + \frac{1}{2}\delta\boldsymbol{a}^T\boldsymbol{H}\delta\boldsymbol{a}. \tag{9}$$

As in previous work[23–25,59] we define a probability distribution $\mathcal{P}$ for fluctuations $\delta\boldsymbol{a}$ around $\boldsymbol{a}_0$. From $\mathcal{P}$ we draw ensembles of perturbed density functionals used for estimating errors on DFT predictions. That the cost function can be assumed to represent the probability of a model given the data is intrinsically a Bayesian idea with no analog in frequentist statistics. We require the mean expectation value of predictions by ensemble models $\boldsymbol{a}' = \boldsymbol{a}_0 + \delta\boldsymbol{a}$ to reproduce the mean prediction error of $\boldsymbol{a}_0$:

$$\sum_j \langle(\delta q_j)^2\rangle_k = \sum_j (\Delta q_j)^2, \tag{10}$$

where $\langle\ldots\rangle_k$ indicates the average over $k \gg 1$ ensemble models. The sums are over j training data while δq_j and Δq_j are prediction errors by $\boldsymbol{a}'$ and $\boldsymbol{a}_0$, respectively. Following Refs. 25 and 59 the probability $\mathcal{P}$ is written

$$\mathcal{P} \propto \exp(-C/T), \tag{11a}$$

$$T = 2C_0/\theta, \tag{11b}$$

where the ensemble temperature T scales the model fluctuations such that Eq. (10) is satisfied. The temperature is in Eq. (11b) expressed in terms of the minimized cost and θ, the effective number of model parameters.

In practical applications of Bayesian error estimation we sample the distribution $\mathcal{P}$. An ensemble matrix $\boldsymbol{\Omega}$ is generated by scaling the inverse Hessian with the ensemble temperature:

$$\boldsymbol{\Omega} = T\boldsymbol{H}^{-1}. \tag{12}$$

Ensemble perturbations $\delta\boldsymbol{a}_k$ are then computed as

$$\delta\boldsymbol{a}_k = \boldsymbol{V}\cdot\boldsymbol{D}\cdot\boldsymbol{u}_k, \tag{13}$$

where matrix $\boldsymbol{V}$ contains the eigenvectors of $\boldsymbol{\Omega}$, matrix $\boldsymbol{D}$ is diagonal and contains the square root of the corresponding eigenvalues, and $\boldsymbol{u}_k$ is a random vector with normal distributed components of zero mean and a spread of 1. The Bayesian error estimate on any DFT prediction from total-energy differences, σ_{BEE}, is then simply related to the variance of $k \gg 1$ non-self-consistent predictions p_k by $\delta\boldsymbol{a}_k$:

$$\sigma_{\text{BEE}} = \sqrt{\text{Var}(\boldsymbol{p})} = \langle\boldsymbol{p}^T\boldsymbol{p}\rangle^{\frac{1}{2}}, \tag{14}$$

where $\boldsymbol{p}$ is a vector of ensemble predictions and the last equality is strictly true for $k \to \infty$, where $\langle\boldsymbol{p}\rangle^2 = 0$.

2. Tikhonov regularization

Smooth exchange enhancement factors are aesthetically pleasing and computationally convenient. Indeed, it was observed in Ref. 25 that smoothness of the enhancement factor is of key importance to obtaining a exchange functional that is transferable to systems not included in the training data. As in Ref. 23 we apply a Tikhonov regularizer R in the cost function Eq. (7) to impose our preference for smooth parametrizations of the MGGA $F_x(s,\alpha)$. This particular regularizer shrinks model coefficients in a "smooth" space $\boldsymbol{b}$,

$$\boldsymbol{b} = \boldsymbol{\Gamma}(\boldsymbol{a} - \boldsymbol{a}_p), \tag{15}$$

where the prior vector $\boldsymbol{a}_p$ is an origo in model space. The prior is thus the resulting fitted model at infinite regularization strength, where all model deviations away from $\boldsymbol{a}_p$ are quenched by the regularizer. The squared Tikhonov matrix $\boldsymbol{\Gamma}^2$ is defined from the overlaps of a scaled Laplacian $\widetilde{\nabla}^2$ of the exchange basis functions $P(t_s, t_\alpha)$,

$$\begin{aligned} \widetilde{\nabla}^2 &= \frac{\partial^2}{\partial t_s^2} + \lambda \frac{\partial^2}{\partial t_\alpha^2}, \\ \boldsymbol{\Gamma}^2_{mnkl} &= \int_{-1}^{1}\int_{-1}^{1} dt_s\, dt_\alpha \widetilde{\nabla}^2 P_{mn} \widetilde{\nabla}^2 P_{kl}, \end{aligned} \tag{16}$$

where λ scales the regularization penalty between polynomials in t_s and t_α. In the present study we have chosen $\lambda = 10^2$, which in numerical tests seems to give a reasonable trade-off between smoothness along s and α. The elements of the Tikhonov matrix grow as the polynomial order of the basis increases, and the exchange model regularization thus preferentially shrinks the more oscillatory components in $F_x(s,\alpha)$. The prior vector $\boldsymbol{a}_p$ in Eq. (15) is chosen such that infinite regularization strength yields $F_x(s,1) = 1$ for all s and half the MS0 exchange along the $F_x(0,\alpha)$ model space direction.

3. Exchange model compromise

The XC model compromises illustrated in Fig. 1 indicate the existence of significant constraints on the performance of semi-local general-purpose density functional approximations: A gain in accuracy on one chemical or materials property is typically associated with a loss of accuracy on a different property. Simultaneously minimizing the prediction errors on several different properties in a transferable manner is therefore a multi-objective (or Pareto) optimization problem. In such Pareto-optimizations, where one cannot a priori infer a strict measure of the relative importance of the individual objectives, there is still one set of solutions that are superior to all other. This is the Pareto set of non-dominated solutions, or the set for which one can not improve one quality without making another quality worse. Among the Pareto-optimal set of solutions one still has a choice in what importance is given to the different qualities. In Ref. 23 a simple but effective approach to this type of problem was developed in the context of density functional fitting, based on minimizing the product of cost functions for the individual training sets including their individual regularizations. The logic underlying this choice is to find a solution among all the Pareto-optimal solutions where the relative improvement of one property leads to a similar relative deterioration of the other properties. The product of cost functions achieves exactly this, if the cost represents the qualities to be optimized. Here we refine that approach by considering a fully regularized cost function for all training data. This corresponds to considering the squared residuals a better measure of quality than the individually minimized (and regularized) cost functions. It is our impression that this improvement offers slightly better fits, and it has the added benefit that the Bayesian interpretation of the statistics is significantly more straightforward, since the functional results from one fit to all data rather than separate fits to each chemical or materials property.

The new starting point for dealing with the exchange model compromise can then be stated as an objective function Φ:

$$\Phi(\boldsymbol{a};\omega) = \Pi_i L_i(\boldsymbol{a}) \times e^{R(\boldsymbol{a};\omega)}, \tag{17}$$

where L_i is the squared-residual loss function for data set i and the exponential a functional form for the prior expectation for the model parameters. Because the logarithm is an injective function, minimizing Φ is equivalent to minimizing $\ln\{\Phi\}$. We can therefore define a regularized cost function K for the exchange model compromise:

$$K(\boldsymbol{a};\omega) = \ln\{\Phi\} = \sum_i \ln\{L_i(\boldsymbol{a})\} + R(\boldsymbol{a};\omega). \tag{18}$$

The minimizing argument vector $\boldsymbol{a}_0(\omega)$, minimizing the objective function K given ω, is a vector that fulfills the zero-gradient condition

$$\frac{\partial K}{\partial \boldsymbol{a}} = 0 = \sum_i \frac{\partial \ln L_i}{\partial \boldsymbol{a}} + \frac{\partial R}{\partial \boldsymbol{a}} = \sum_i \frac{1}{L_i}\frac{\partial L_i}{\partial \boldsymbol{a}} + \frac{\partial R}{\partial \boldsymbol{a}}. \tag{19}$$

If K had been quadratic and positive definite, the existence of only a single solution vector $\boldsymbol{a}_0(\omega)$ would have been certain. This, however, does not appear to be a significant problem, at least not with the data sets we have fitted in the present study. Since the loss function, $L_i(\boldsymbol{a})$, and the regularizer, $R(\boldsymbol{a};\omega)$, are both quadratic in $\boldsymbol{a}$, the zero-gradient condition above is very close to representing a traditional least-squares minimization problem, and we solve it by iterative least-squares minimization by casting it on the form of Eq. (7):

$$\begin{aligned} \widetilde{K}(\boldsymbol{a};\omega,\boldsymbol{a}_*) &= \widetilde{L}(\boldsymbol{a};\boldsymbol{a}_*) + R(\boldsymbol{a};\omega) \\ &= \sum_i \frac{L_i(\boldsymbol{a})}{L_i(\boldsymbol{a}_*)} + R(\boldsymbol{a};\omega) \\ &= \tilde{\boldsymbol{q}}^T \tilde{\boldsymbol{q}} + \omega^2 \boldsymbol{b}^T \boldsymbol{b}, \end{aligned} \tag{20}$$

with least-squares solution $\tilde{\boldsymbol{a}}$. This solution is then inserted for $\boldsymbol{a}_*$ iteratively, and convergence is reached in very few steps, when $\tilde{\boldsymbol{a}} = \boldsymbol{a}_*$. In that case, the model-compromise cost Eq. (20) reduces to $\widetilde{K} = N_D + \omega^2 \boldsymbol{b}^T \boldsymbol{b}$, with N_D the number of training data sets.

The concept of an effective number of model parameters, as defined in Eq. (8), applies equally well to $\widetilde{K}$, as does the definition of the Bayesian ensemble matrix in Eq. (12). Only the model complexity $\hat{\theta}$, corresponding to the globally optimum exchange model $\hat{\boldsymbol{a}}_0$, remains to be determined. This model should constitute a suitable trade-off between model bias and variance such that it generalizes well to properties outside the training sets.[58] We here apply a clustered leave-one-out cross validation estimator of the generalization error, Δ^2:

$$\Delta^2(\omega) = \frac{1}{N_D} \sum_{i=1}^{N_D} L_i(\tilde{\boldsymbol{a}}_i(\omega)), \tag{21}$$

where training set i has been excluded from $\widetilde{K}$ when determining $\tilde{\boldsymbol{a}}_i$.

In summary, we thus determine the optimal simultaneous fit to all training data in the protocol, $\hat{\boldsymbol{a}}_0$, by identifying the regularization strength $\hat{\omega}_0$ that minimizes the generalization error Δ^2. The corresponding exchange model complexity is $\hat{\theta}$, and Bayesian error estimates on materials property predictions by $\hat{\boldsymbol{a}}_0$ are obtained following Eqs. (11b)–(14).

VI. RESULTS

A. mBEEF density functional and BEE ensemble

Figure 2 shows a range of meta-GGA exchange enhancement factors obtained by minimizing Eq. (20) for increasing model complexities, i.e., for decreasing ω. The enhancement factors are neatly smooth along s (top panel) and α (bottom panel) for small θ, but develop increasingly non-smooth features when the exchange models are allowed to become more complex, particularly for $\theta > 12$. The optimum trade-off between performance and transferability, as determined by minimizing Δ^2, we find at $\hat{\theta} = 8.8$. This model we henceforth denote mBEEF exchange. It is indicated by full black lines in Fig. 2. Note that the full mBEEF exchange-correlation functional uses PBEsol correlation, see Eq. (6), and that mBEEF exchange does not conform to the formal UEG limit. This appears to be a quite general feature of semi-local DFAs optimized for chemistry.[23,25,41] Consequently, the mBEEF enhancement of LDA exchange for a UEG-like electronic structure at $(s, \alpha) = (0, 1)$ is $F_x(0, 1) = 1.037$, while for rapidly varying densities $F_x(\infty, 1) = 1.145$. The latter is a significantly lower exchange enhancement in the large gradient/small density regime than for most semi-local functionals.

The mBEEF error estimation exchange ensemble is illustrated in Fig. 3. Note how constrained the ensemble is around $(s, \alpha) = (0, 1)$, and that it clearly straddles the UEG limit in this point. The ensemble models spread out significantly for $(s, \alpha) > (2, 2)$, indicating that the functional form of the mBEEF $F_x(s, \alpha)$ is less constrained in this region of the MGGA electronic-structure parameter space. We reported similar findings in Ref. 23 for the large-s regime of the BEEF-vdW ensemble. The fact that the ensemble is very broad for large reduced density gradients suggests that the decay of the mBEEF $F_x(s, 1)$ towards 1.145 for $s \to \infty$ is not imposed by the training data sets. Rather, the training data offers very little electronic-structure information for $(s, \alpha) > (2, 2)$, and the exchange model in this region therefore becomes

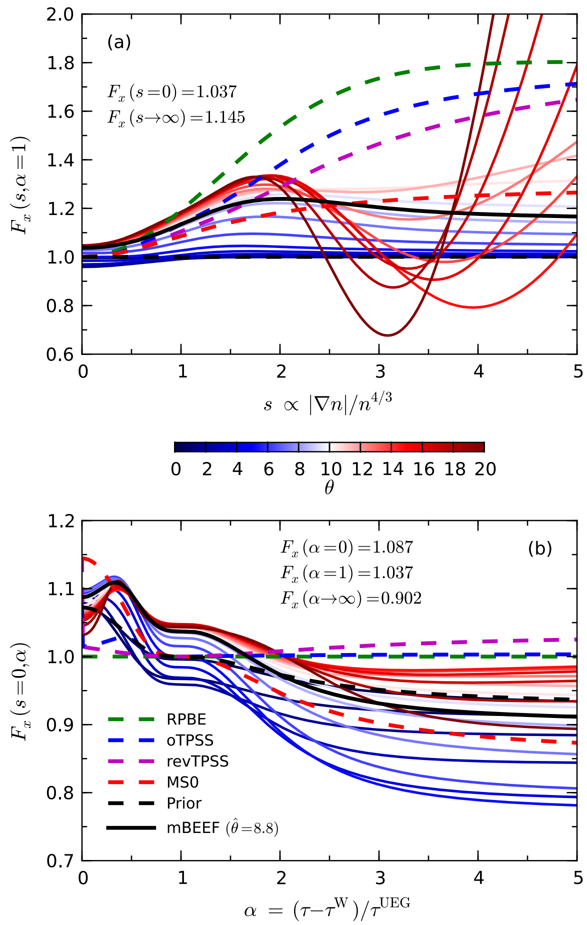


FIG. 2. Model-compromise optimized mBEEF type exchange enhancement factors for increasing number of effective parameters $\theta \in [0, 20]$. Full blue lines indicate $\theta < 10$, while full red lines indicate $\theta > 10$. Full black lines illustrate the chosen mBEEF $F_x(s, \alpha)$. Standard GGA and MGGA exchange functionals are illustrated by dashed lines along with the prior model. a) Projections along s for $\alpha = 1$. Note that $\alpha = 1$ for a uniform electron gas, and that for this value of the reduced KED the MGGA $F_x(s, \alpha)$ is equivalent to a GGA exchange enhancement factor. b) Projections along α for $s = 0$. All but the most constrained mBEEF type exchange functionals have a curved feature betwen the single-electron limit ($\alpha = 0$) and the UEG region ($\alpha \approx 1$).

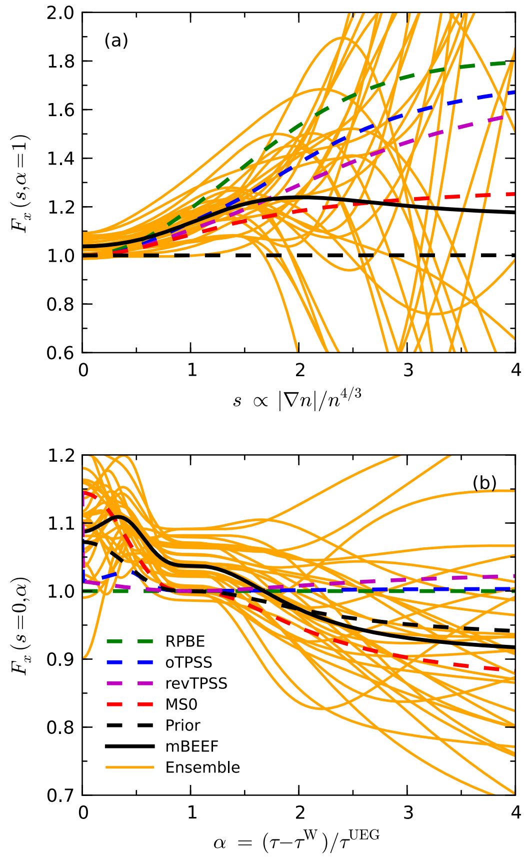


FIG. 3. Bayesian ensemble of exchange models (yellow) around the mBEEF (solid black). Standard GGA and MGGA exchange functionals are illustrated by dashed lines along with the prior model. a) Projections along s for $\alpha = 1$. b) Projections along α for $s = 0$. The ensemble is rather constrained around the HEG limit $(s, \alpha) = (0, 1)$ in both panels, but spreads out significantly away from this region of XC mocel space.

strongly dominated by the prior model $\boldsymbol{a}_p$.

The mBEEF exchange expansion vector and error estimation ensemble matrix are freely available online[60] for easy numerical implementation in DFT codes already implementing meta-GGA functionals.

B. Benchmark

Figure 4 summarizes a broad benchmark of some popular or recent GGA and MGGA density functionals in addition to the mBEEF and BEEF-vdW functionals. All data is obtained from selfconsistent DFT calculations. The bars indicate the logarithms of scaled mean-absolute errors on the five training data sets applied in mBEEF fitting. The mBEEF exchange model compromise appears quite reasonable, as one would expect, since the functional was trained on this data: The MAE is among the three lowest for all five properties and presents a considerable improvement over the BEEF-vdW in predicting the lattice constants and cohesive energies of bulk solids, while not compromising the good description of the adsorbate–surface bond strengths in CE27a, which is almost on the level of the RPBE functional.

We further underline this point by showing on a logarithmic scale in Fig. 5 the product of the corresponding RMS errors relative to that of mBEEF. It is apparent that mBEEF on the training sets simultaneously achieve very acceptable predictions within the five classes of chemical and materials properties.

C. Transferability

We shall now assess the mBEEF transferability by considering quantities outside the training data. Table I compares error statistics on the MB08-165 decomposition energies of artificial molecules, the BM32 bulk moduli, the SE30 surface energies, and 26 of the 27 binding energies of neutral and charged water clusters in the WATER27 benchmark set. The mBEEF functional appears to generalize reasonably well to prediction of properties not explicitly included in the training sets used to generate it. The decomposition energies and bulk moduli are on average predicted with only a limited systematic bias. The surface energies are on average underestimated. For this property mBEEF performs better than PBE and is nearly on par with MS0, but does not attain the accuracy of the TPSS-class functionals. As observed in Fig. 1, this may be due to mBEEF's focus on performing well for chemisorption energies. Interestingly, the water cluster binding energies are surprisingly well captured by mBEEF even though systems with significant noncovalent interactions were not included in the training data. Contrary to the two TPSS-type MGGAs, PBE and MS0 also appear to describe this sort of hydrogen bonding well. Similar findings were reported in Ref. 40. Another MS0-type MGGA was in Refs. 61 and 62 also successfully applied to systems with weak bonding. All together this suggests that the high accuracy of mBEEF for hydrogen bonding may to some extend be due to the use of a MS0-based form of the α-dependence in the exchange model space. We therefore concur with the hypothesis of Perdew *et al.*[62] that future high-performance van der Waals (vdW) density functionals might benefit greatly from optimized MS0-based exchange.

We note in passing that mBEEF also correctly predicts the sequence of relative stabilities of the 4 isomers of the water hexamer included in the WATER27 set. Moreover, the MAE over the 4 isomers is less than 1 kcal/mol. Most semi-local DFAs agree much worse with benchmark quantum chemical calculations on these systems. According to literature, it usually takes highly specialized exchange-correlation functionals optimized for hydrogen bonding[63] or dedicated vdW functionals[44] to get the energetic ordering of water hexamers right. MS0 correctly predicts the ordering when the benchmark (B3LYP) structures

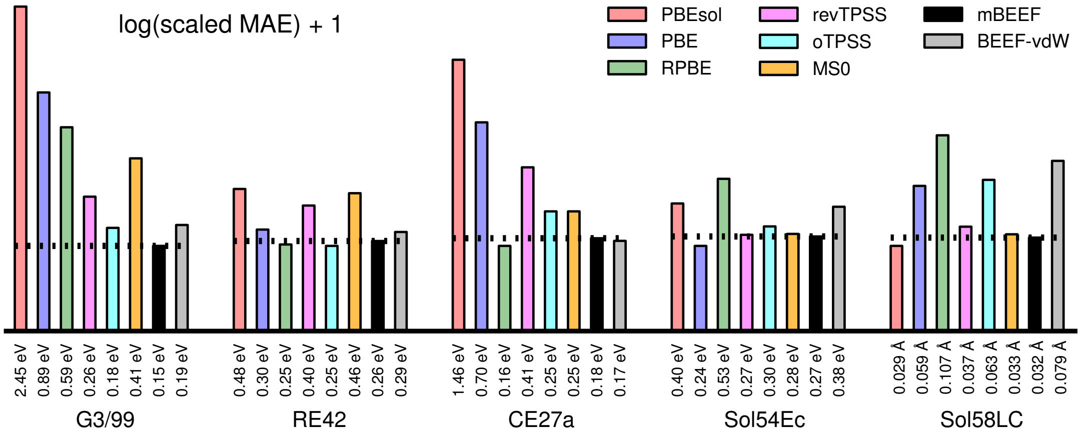


FIG. 4. Benchmark of mBEEF against popular or recent GGA and MGGA density functionals in terms of mean-absolute errors (MAEs) on predicting the chemical and materials properties represented by the 5 data sets applied in mBEEF training. BEEF-vdW is also included. Note that each bar is normlized with the smallest one and plotted on a logarithmic scale for reasons of clarity. Horizontal black dash-dotted lines indicate the mBEEF level, which is among the 3 lowest for all five properties.

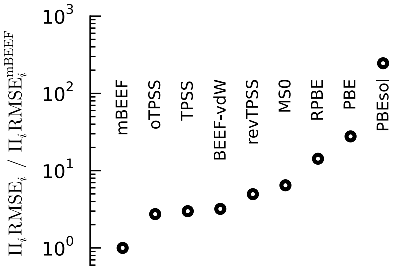


FIG. 5. Relative products of RMS errors on the five mBEEF training sets. Note the logarithmic scale and the clustering of the density functional approximations into GGAs and MGGAs+BEEF-vdW. This is a direct consequence of improved possibilities for the XC model compromise.

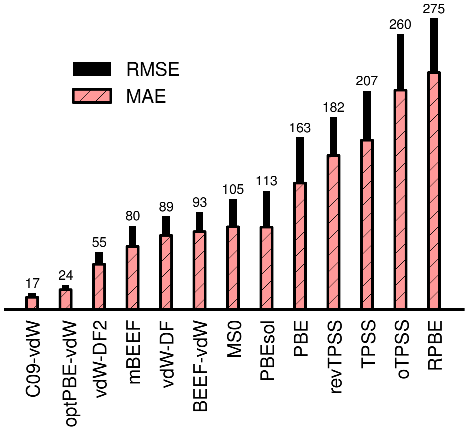


FIG. 6. Comparison of various density functional prediction errors on the S22 non-covalent benchmark set. Numerical values for the RMS errors are indicated in units of meV above each set of bars.

are used, but fails upon structural relaxation, which destabilizes the 'prism' isomer by 1.3 kcal/mol. This is not the case with mBEEF, where relaxation leads to near-isoenergetics for the 'prism' and 'cage' isomers.

We further highlight the interesting finding above that mBEEF performs surprisingly well for non-covalently bonded systems by considering the S22 quantum chemical benchmark set[64,65] for non-covalently bonded complexes. This data set exhibits hydrogen bonding as well as van der Waals dispersion. Figure 6 shows error statistics for several GGA, MGGA, and vdW-DF type density functionals in reproducing the S22 binding energies. Semi-local DFAs do not contain the physics needed to reliably capture long-ranged dispersion interactions. It is therefore no surprise that the largest prediction errors in Fig. 6 are found for GGAs and MGGAs, while vdW functionals with explicitly non-local correlation are better suited for this. Though mBEEF is a semi-local functional and is not explicitly designed to capture dispersion interactions, its performance on S22 is good. Notably, mBEEF for example seems to on average outperform the significantly more expensive vdW-DF functional. We would expect even better performance on the S22 benchmark if a suitable non-local correlation term[42,66] was added to the mBEEF model space.

D. Bayesian error estimates: The CO puzzle

Finally, let us consider an example of applying the BEE approach to error estimation in DFT. We choose a prototypical surface chemical problem: Predicting the site preference of molecular CO adsorption on close-packed surfaces of late transition metals. Most semi-local density functionals fail to correctly predict the most stable adsorption site over several such metals. This 'CO puz-

TABLE I. Error statistics for different density functionals in predicting various chemical and materials properties not included in mBEEF training. Computed statistics are mean error (ME) or mean relative error (MRE) and their absolute counterparts.

	LSDA	PBEsol	PBE	RPBE	revTPSS	oTPSS	MS0	mBEEF	BEEF-vdW
	MB08-165 decomposition energies of artificial molecules[a] *(kcal/mol)*								
ME	15.4	7.6	1.4	−4.9	−7.3	−2.3	−11.0	0.1	−2.0
MAE	19.9	12.7	9.0	11.3	13.2	6.8	18.4	8.1	12.2
	BM32 bulk moduli[b] *(%)*								
MRE	5.7	−3.0	−10.7	−17.9	−3.2	−7.3	−0.8	−0.7	−12.8
MARE	7.9	5.1	10.9	17.9	6.8	8.6	5.5	7.1	14.7
	SE30 surface energies[c] *(%)*								
MRE	−7	−13	−26	−35	−6	−11	−18	−22	−21
MARE	14	17	26	35	12	15	20	23	23
	WATER26 binding energies of neutral and charged water clusters[d] *(%)*								
MRE	47.5	17.3	2.7	−18.8	−7.8	−13.2	−2.5	2.3	−12.5
MARE	47.5	17.3	3.6	18.9	7.8	13.6	2.7	2.7	12.5

[a] Quantum chemical benchmark from Ref. 67.

[b] 32 experimental bulk moduli from Refs. 68 and 69, all corrected for thermal contributions and zero-point phonon effects.

[c] 30 experimental surface energies from Ref. 70.

[d] Quantum chemical benchmark from Ref. 71. This set was in Ref. 43 named WATER27, but we exlude here the last benchmark data point since it is a conformational energy difference rather than a binding energy.

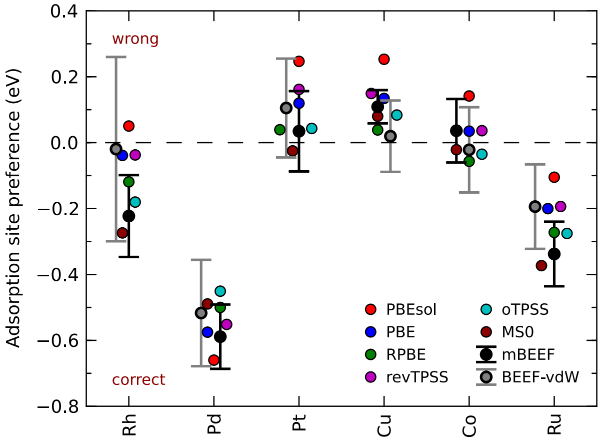


FIG. 7. Site preference ΔE for CO adsorption on (111) surfaces of Rh, Pd, Pt, and Cu and (0001) surfaces of Co and Ru at 0.25 monolayer coverage. Error bars on mBEEF and BEEF-vdW predictions indicate Bayesian error estimates.

zle' is a standing issue in computational surface chemistry, and a large number of studies have been devoted to elucidating its origin and its possible solutions, see for example Refs. 72–78.

Figure 7 shows calculated adsorption energy differences ΔE between the experimentally most stable CO adsorption site and less stable sites among the hollow and atop sites on close-packed facets of Rh, Pd, Pt, Cu, Co, and Ru, such that $\Delta E < 0$ eV corresponds to a correct theoretical prediction of the most stable of the two sites. Predictions made using a range of DFAs are indicated with different colors in the figure. The experimentally observed preference at low temperature and coverage is for the 1-fold coordinated atop site on all of the considered surfaces except Pd(111), on which the 3-fold coordinated fcc site is found to be energetically most favorable.

Bayesian error estimates σ_{BEE} are shown for mBEEF and BEEF-vdW calculations. Most GGAs and MGGAs correctly predict $\Delta E < 0$ eV on Rh(111), Pd(111), and Ru(0001), while on Pt(111), Cu(111), and Co(0001) the theoretical predictions are scattered around or just above $\Delta E = 0$ eV. The mBEEF σ_{BEE} values provide very reasonable estimates of the spread of predictions by different GGA or MGGA density functionals. In particular, the BEEs indicate that calculated adsorption site preferences for CO on Pt(111) and Co(0001) should not be considered indisputable, but may well change depending on the choice of exchange-correlation functional. In some sense Fig. 7 shows that such sensitivities of scientific conclusions are also found if we meticulously compute each ΔE using a wide range of different DFAs. However, Bayesian error estimation ensembles provide a quantitative and computationally inexpensive approach to such analysis.

VII. CONCLUSIONS

Broadly applicable semi-local density functionals must somehow be designed with the exchange-correlation model compromise in mind. The XC model selection procedure in the Bayesian error estimation functional framework effectively addresses this multi-objective optimization problem. We here used it to develop the mBEEF exchange-correlation functional, and argue that this can be considered a very reasonable general-purpose meta-GGA density functional. It delivers highly accurate predictions of a wide range of different properties in materials physics and chemistry, and we expect mBEEF to be particularly well suited for computational studies in surface science, including catalysis. A Bayesian ensemble for error estimation in DFT is an intrinsic feature of the BEEF-class of density functionals. The ensemble is defined in terms of XC model fluctuations, and we have illustrated the application of error estimation by considering a prototypical surface chemical problem. The mBEEF ensemble error estimates correctly indicate that one can not conclusively determine the site-preference of CO adsorption on a number of late transition metal surfaces. A DFT user may traditionally try to get some idea about the sensitivity of calculated quantities on the choice of density functional approximation by tediously applying various established functionals to the same problem. The BEE provides a more structured approach to such analysis via systematic but computationally inexpensive computations of non-self-consistent XC energy perturbations. We expect this approach to quantitative error estimation of correlated errors to become a useful and very general tool for validating scientific conclusions based on DFT in computational materials science and chemistry. Finally, we find that the mBEEF functional captures the strength of hydrogen bonding and van der Waals bonding reasonably well, even though it was not explicitly designed for this. This suggests that mBEEF may be a very appropriate starting point for a meta-GGA exchange-correlation functional explicitly including non-local van der Waals correlation to accurately account for long-range dispersion interactions. This will be the topic of future extensions of this work.

VIII. ACKNOWLEDGMENTS

We acknowledge financial support to SUNCAT from the Chemical Sciences, Geosciences and Biosciences Division of the U.S. Department of Energy's Office of Basic Energy Science. We also wish to thank the Danish Center for Scientific Computing for computational resources.

[1] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).

[2] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).

[3] K. Burke, J. Chem. Phys. **136**, 150901 (2012).

[4] G. H. Jóhannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen, and J. K. Nørskov, Phys. Rev. Lett. **88**, 255506 (2002).

[5] T. Bligaard, G. H. Jóhannesson, A. V. Ruban, H. L. Skriver, K. W. Jacobsen, and J. K. Nørskov, Appl. Phys. Lett. **83**, 4527 (2003).

[6] D. Morgan, G. Ceder, and S. Curtarolo, Meas. Sci. Technol. **16**, 296 (2005).

[7] W. Setyawan, R. M. Gaume, S. Lam, R. S. Feigelson, and S. Curtarolo, ACS Comb. Sci. **13**, 382 (2011).

[8] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, Comput. Mat. Sci. **50**, 2295 (2011).

[9] I. E. Castelli, T. Olsen, S. Datta, D. D. Landis, S. Dahl, K. S. Thygesen, and K. W. Jacobsen, Energy Environ. Sci. **5**, 5814 (2012).

[10] G. Hautier, A. Miglio, G. Ceder, G.-M. Rignanese, and X. Gonze, Nat. Commun. **4**, 2292 (2013).

[11] P. Strasser, Q. Fan, M. Devenney, W. H. Weinberg, P. Liu, and J. K. Nørskov, J. Phys. Chem. B **107**, 11013 (2003).

[12] J. Greeley, T. F. Jaramillo, J. Bonde, I. Chorkendorff, and J. K. Nørskov, Nat. Mater. **5**, 909 (2006).

[13] M. P. Andersson, T. Bligaard, A. Kustov, K. E. Larsen, J. Greeley, T. Johannessen, C. H. Christensen, and J. K. Nørskov, J. Catal. **239**, 501 (2006).

[14] F. Studt, F. Abild-Pedersen, T. Bligaard, R. Z. Sørensen, C. H. Christensen, and J. K. Nørskov, Science **320**, 1320 (2008).

[15] J. K. Nørskov, T. Bligaard, J. Rossmeisl, and C. H. Christensen, Nat. Chem. **1**, 37 (2009).

[16] G. Hautier, A. Jain, H. Chen, C. Moore, S. P. Ong, and G. Ceder, J. Mater. Chem. **21**, 17147 (2011).

[17] W. Kohn, A. D. Becke, and R. G. Parr, J. Phys. Chem. **100**, 12974 (1996).

[18] J. P. Perdew, A. Ruzsinszky, J. Tao, V. N. Staroverov, G. E. Scuseria, and G. I. Csonka, J. Chem. Phys. **123**, 062201 (2005).

[19] Y. Zhao and D. G. Truhlar, J. Chem. Phys. **128**, 184109 (2008).

[20] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, Phys. Rev. Lett. **100**, 136406 (2008).

[21] R. Peverati and D. G. Truhlar, J. Chem. Phys. **136**, 134704 (2012).

[22] A. Ruzsinszky, G. I. Csonka, and G. E. Scuseria, J. Chem. Theory Comput. **5**, 763 (2009).

[23] J. Wellendorff, K. T. Lundgaard, A. Møgelhøj, V. Petzold, D. D. Landis, J. K. Nørskov, T. Bligaard, and K. W. Jacobsen, Phys. Rev. B **85**, 235149 (2012).

[24] J. J. Mortensen, K. Kaasbjerg, S. L. Frederiksen, J. K. Nørskov, J. P. Sethna, and K. W. Jacobsen, Phys. Rev. Lett. **95**, 216401 (2005).

[25] V. Petzold, T. Bligaard, and K. W. Jacobsen, Top. Catal. **55**, 402 (2012).

[26] R. Y. Brogaard, P. G. Moses, and J. K. Nørskov, Catal. Lett. **142**, 1057 (2012).

[27] M. Dell'Angela, T. Anniyev, M. Beye, R. Coffee, A. Föhlisch, J. Gladh, T. Katayama, S. Kaya, O. Krupin, J. LaRue, A. Møgelhøj, D. Nordlund, J. K. Nørskov, H. Öberg, H. Ogasawara, H. Öström, L. G. M. Pettersson, W. F. Schlotter, J. A. Sellberg, F. Sorgenfrei, J. J. Turner, M. Wolf, W. Wurth, and A. Nilsson, Science **339**, 1302 (2013).

[28] F. Studt, F. Abild-Pedersen, J. B. Varley, and J. K. Nørskov, Catal. Lett. **143**, 71 (2013).

[29] R. Y. Brogaard, B. M. Weckhuysen, and J. K. Nørskov, J. Catal. **300**, 235 (2013).

[30] B. Hammer, L. B. Hansen, and J. K. Nørskov, Phys. Rev. B **59**, 7413 (1999).

[31] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[32] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, Phys. Rev. Lett. **100**, 136406 (2008).

[33] P. Haas, F. Tran, P. Blaha, and K. Schwarz, Phys. Rev. B **83**, 205117 (2011).

[34] A. D. Becke, J. Chem. Phys. **109**, 2092 (1998).

[35] J. P. Perdew, S. Kurth, A. Zupan, and P. Blaha, Phys. Rev. Lett. **82**, 2544 (1999).

[36] A. D. Boese and N. C. Handy, J. Chem. Phys. **116**, 9559 (2002).

[37] J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria, Phys. Rev. Lett. **91**, 146401 (2003).

[38] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, L. A. Constantin, and J. Sun, Phys. Rev. Lett. **103**, 026403 (2009).

[39] R. Peverati and D. G. Truhlar, J. Phys. Chem. Lett. **3**, 117 (2012).

[40] J. Sun, B. Xiao, and A. Ruzsinszky, J. Chem. Phys. **137**, 051101 (2012).

[41] A. D. Boese and N. C. Handy, J. Chem. Phys. **114**, 5497 (2001).

[42] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist, Phys. Rev. Lett. **92**, 246401 (2004).

[43] L. Goerigk and S. Grimme, J. Chem. Theory Comput. **6**, 107 (2010).

[44] J. Klimes, D. R. Bowler, and A. Michaelides, J. Phys.: Condens. Matter **22**, 022201 (2010).

[45] V. R. Cooper, Phys. Rev. B **81**, 161104(R) (2010).

[46] Linear relationships between DFT-predicted chemisorption energies and surface energies have appeared in recent literature for the particular cases of CO on Pt(111) and Rh(111), see Refs. 75, 77, and 79. However, bivariate analyses of DFT prediction errors for surface chemistry and stability has, to the authors' knowledge, not previously been considered on such firm statistical footing as in Fig. 1.

[47] B. G. Janesko, V. Barone, and E. N. Brothers, J. Chem. Theory Comput. **9**, 4853 (2013).

[48] L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople, J. Chem. Phys. **112**, 7374 (2000).

[49] The G3/99 standardization factor is $1/(N_\mathrm{a}-1)$, where N_a is the number of atoms in each molecule. For RE42 the factor is $1/(N_\mathrm{r}-N_\mathrm{p}+1)$, where N_r and N_p are the number of reactants and products in each reaction, respectively.

[50] The CE27a experimental chemisorption energies are essentially from Ref. 23, but are here referenced to free atoms rather than gas-phase adsorbates. The Sol54Ec and Sol58LC datasets both derive from related sets applied in Ref. 23, containing experimental cohesive energies and lattice constants of monoatomic fcc, bcc, and diamond structured crystals corrected for thermal and vibrational contributions. We here also include diatomic rocksalt, cesiumchloride, and zincblende crystals, and update the lattice constant phonon corrections with recent *ab initio* values from Ref. 80. Note that Sol54Ec and Sol58LC contain data for Pb(fcc) which is excluded in model training. The cohesive energy of lead appears a very significant outlier in model training, and is known to be severely overestimated in DFT, see Ref. 81 and the Supplemental Material for Ref. 23.

[51] J. J. Mortensen, L. B. Hansen, and K. W. Jacobsen, Phys. Rev. B **71**, 035109 (2005).

[52] J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Duak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. A. Hansen, H. H. Kristoffersen, M. Kuisma, A. H. Larsen, L. Lehtovaara, M. Ljungberg, O. Lopez-Acevedo, P. G. Moses, J. Ojanen, T. Olsen, V. Petzold, N. A. Romero, J. Stausholm-Møller, M. Strange, G. A. Tritsaris, M. Vanin, M. Walter, B. Hammer, H. Häkkinen, G. K. H. Madsen, R. M. Nieminen, J. K. Nørskov, M. Puska, T. T. Rantala, J. Schiøtz, K. S. Thygesen, and K. W. Jacobsen, J. Phys.: Condens. Matter **22**, 253202 (2010).

[53] P. E. Blöchl, Phys. Rev. B **50**, 17953 (1994).

[54] S. R. Bahn and K. W. Jacobsen, Comput. Sci. Eng. **4**, 56 (2002).

[55] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).

[56] A. B. Alchagirov, J. P. Perdew, J. C. Boettger, R. C. Albers, and C. Fiolhais, Phys. Rev. B **63**, 224115 (2001).

[57] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. (Springer, 2006).

[58] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, 2009).

[59] S. L. Frederiksen, K. W. Jacobsen, K. S. Brown, and J. P. Sethna, Phys. Rev. Lett. **93**, 165501 (2004).

[60] See supplemental material at [URL] for the 64 mBEEF exchange expansion coefficients and the 64×64 mBEEF error estimation ensemble matrix.

[61] J. Sun, R. Haunschild, B. Xiao, I. W. Bulik, G. E. Scuseria, and J. P. Perdew, J. Chem. Phys. **138**, 044113 (2013).

[62] J. Sun, B. Xiao, Y. Fang, R. Haunschild, P. Hao, A. Ruzsinszky, G. I. Csonka, G. E. Scuseria, and J. P. Perdew, Phys. Rev. Lett. **111**, 106401 (2013).

[63] E. E. Dahlke, R. M. Olson, H. R. Leverentz, and D. G. Truhlar, J. Phys. Chem. A **112**, 3976 (2008).

[64] P. Jurecka, J. Sponer, J. Cerny, and P. Hobza, Phys. Chem. Chem. Phys. **8**, 1985 (2006).

[65] T. Takatani, E. G. Hohenstein, M. Malagoli, M. S. Marshall, and C. D. Sherrill, J. Chem. Phys. **132**, 144104 (2010).

[66] O. A. Vydrov and T. Van Voorhis, J. Chem. Phys. **133**, 244103 (2010).

[67] M. Korth and S. Grimme, J. Chem. Theory Comput. **5**, 993 (2009).

[68] L. Schimka, J. Harl, and G. Kresse, J. Chem. Phys. **134**, 024116 (2011).

[69] G. I. Csonka, J. P. Perdew, A. Ruzsinszky, P. H. T. Philipsen, S. Lebègue, J. Paier, O. A. Vydrov, and J. G. Ángyán, Phys. Rev. B **79**, 155107 (2009).

[70] L. Vitos, A. V. Ruban, H. L. Skriver, and J. Kollár, Surf. Sci. **411**, 186 (1998).

[71] V. S. Bryantsev, M. S. Diallo, A. C. T. van Duin, and W. A. Goddard III, J. Chem. Theory Comput. **5**, 1016 (2009).

[72] P. J. Feibelman, B. Hammer, J. K. Nørskov, F. Wagner, M. Scheffler, R. Stumpf, R. Watwe, and R. Dumesic, J. Phys. Chem. B **105**, 4018 (2001).

[73] M. Gajdos, A. Eichler, and J. Hafner, J. Phys.: Condens. Matter **16**, 1141 (2004).

[74] F. Abild-Pedersen and M. P. Andersson, Surf. Sci. **601**, 1747 (2007).

[75] A. Stroppa and G. Kresse, New J. Phys. **10**, 063020 (2008).

[76] X. Ren, P. Rinke, and M. Scheffler, Phys. Rev. B **80**, 045402 (2009).

[77] L. Schimka, J. Harl, A. Stroppa, A. Grüneis, M. Marsman, F. Mittendorfer, and G. Kresse, Nat. Mater. **9**, 741 (2010).

[78] P. Lazić, M. Alaei, N. Atodiresei, V. Caciuc, R. Brako, and S. Blügel, Phys. Rev. B **81**, 045401 (2010).

[79] J. Sun, M. Marsman, A. Ruszinszky, G. Kresse, and J. P. Perdew, Phys. Rev. B **83**, 121410(R) (2011).

[80] P. Hao, Y. Fang, J. Sun, G. I. Csonka, P. H. T. Philipsen, and J. P. Perdew, Phys. Rev. B **85**, 014111 (2012).

[81] D. Yu and M. Scheffler, Phys. Rev. B **70**, 155417 (2004).