MULTIVARIATE LINEAR FREE ENERGY RELATIONSHIPS IN CATALYSIS AND DRUG BINDING

by

ANDREA NICOLE BOOTSMA

(Under the Direction of Steven Wheeler)

ABSTRACT

While linear free energy relationships (LFERs) have been used for decades to relate electronic and steric features of a substituent to the reaction rate, and thus infer something about the reaction mechanism they have recently become more popular as tools for catalyst and ligand design. In these cases, multi-parameter fitting is used to create a function that predicts an experimental outcome - such as reaction rate, selectivity, binding affinity, or inhibitory concentration - based on computationally determined descriptors of the ligand or catalyst. However, these fits are often complex and difficult to interpret. I have proposed and developed a series of multivariate LFERs that use purely computational data. This allows us to isolate features of the complex chemical system under consideration, and thus achieve both quantitative prediction of interaction energetics and clear interpretability, providing chemical insight into the underlying interactions that influence the observed trends. I have applied this methodology to both the areas of catalyst design and drug optimization. In the area of catalyst design I was able to propose a new catalyst for a BINOL-catalyzed asymmetric conjugate addition which has since been experimentally shown to have improved yield.

In the area of drug design and optimization I have developed new descriptors of heterocycle electrostatics that have been successfully used to describe both traditional and non-traditional stacking interactions that, while observed in drug binding sites, had previously not been well understood or fully taken advantage of. Further, I have developed methodology for the rapid evaluation of these descriptors from atom connectivity information such as a SMILES string, eliminating the need for expensive quantum mechanical computations to make accurate predictions.

INDEX WORDS: computational chemistry, density functional theory, catalyst design, noncovalent interactions, molecular descriptors, multivariate linear regression

MULTIVARIATE LINEAR FREE ENERGY RELATIONSHIPS IN CATALYSIS AND DRUG BINDING

by

ANDREA NICOLE BOOTSMA

B.S., Calvin College, 2015

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2019

MULTIVARIATE LINEAR FREE ENERGY RELATIONSHIPS IN CATALYSIS AND DRUG BINDING

by

ANDREA NICOLE BOOTSMA

Major Professor:	Steven E. Wheeler
Committee:	Henry F. Schaefer
	Robert Phillips

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2019

DEDICATION

To Dr. Carolyn E. Anderson, my undergraduate research advisor. You helped me find the path that lead me to this point.

ACKNOWLEDGEMENTS

I am grateful to so many people who have helped me along the journey to completing my PhD. I do not exist in a vacuum and I know I could have never accomplished this on my own. Firstly, I thank my research advisor Dr. Steven E. Wheeler, who has generously guided me through these four years as a graduate student, answering my questions about research, my career and more. I also thank the other members of the Wheeler group, Drew, Stephen, SharathChandra, Yanfei, Rajat, Bryan, Victoria, Laura, and Tony, who have been a pleasure to work alongside, and consistently offer good advice, and encouragement. I also thank the members of my committee Dr. Henry F. Schaefer and Dr. Robert Phillips, for their commitment to my success and support along the way.

My family has been an irreplaceable support to me over these four years, and in everything that led me here. So, thank you to Mom & Dad, Paul, Aleca and Scarlett, for encouraging me through stress and tears and moves and more, for being proud of me, and for reminding me of life outside of graduate school. Thanks also to Pake & Beppe, Grandma & Grandpa and all my aunts, uncles, cousins and beyond, I have so appreciated your prayers and your interest in the work that I do. I am also grateful to my church families at both Fellowship Church in College Station, TX and Resurrection Presbyterian Church in Athens, GA who have been praying for me and encouraging me throughout this journey. And finally, I want to thank my Heavenly Father, without whom none of this would be possible, and who sustains my life every single day.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Motivation for Studying Organic Systems using Computational Chemistry

Computational chemistry uses physical principles to compute the properties of molecules such as the structure, relative energy, free energy and electrostatic potential. The properties can in turn be used to explain observed chemical phenomena, and to make predictions about the behavior of systems for which experimental data is not available. This feature is particularly valuable in cases where experiments are dangerous, expensive, or time consuming to perform, such as is often the case in the design of selective catalysts and drug candidates. In these cases, it is desirable to avoid the synthesis of ineffective catalysts or ligands, and thus computational methods for determining the top candidates are appealing. However, useful predictions require accurate computation of the molecular properties, which in turn often requires the use of methods that are computationally expensive. As computational cost scales steeply with the size of the system under investigation, this makes large scale computational analysis of relevant systems prohibitively expensive and hamstrings efforts toward computationally driven molecular design.

In this case Linear Free Energy Relationships (LFERs) can be incredibly useful. Initially developed by Hammett in order to measure electronic effects in the transition state in 1937,[1] LFERs are powerful tools capable of providing both predictive power and mechanistic understanding. They capture underlying trends in the observed behavior, and

allow us to make accurate predictions about unknown behavior based on more easily accessible steric and electronic properties of the molecule. Often, for the complex task of catalyst or ligand design a single descriptor is inadequate to describe the factors influencing a given process or property. In these cases, multi-parameter fitting is used to create a function that predicts an experimental outcome—such as reaction rate, selectivity, binding affinity, or inhibitory concentration—based on computationally determined descriptors of the ligand or catalyst. These relationships can reveal the ways that different properties may work in concert or conflict, and inform the design process whereby features are changed simultaneously (*i.e.* choosing a new substituent that is both sterically and electronically distinct). In the best case application of these techniques, the combination of predictive power and interpretability means that these relationships can be used for both design and understanding.

While these techniques have been traditionally used to relate experimentally or computationally described molecular descriptors to experimental data regarding rates, binding constants, or selectivities, I have extended their use to understanding and predicting purely computational results such as transition state barriers and computed binding energies. This allows us to isolate features of the complex chemical system under consideration, and thus achieve both quantitative prediction of interaction energetics and clear interpretability, providing chemical insight into the underlying interactions that influence the observed trends. I have applied this methodology to both catalyst and drug design. In the area of catalyst design, I was able to propose a new catalyst for a BINOL-catalyzed asymmetric conjugate addition which has since been experimentally shown to have improved yield and selectivity. In the area of drug design and optimization, I have

developed new descriptors of heterocycle electrostatics that have been successfully used to both predict and understand traditional and non-traditional stacking interactions between heterocycles and functional groups appearing in drug binding sites.

Motivation for Studying Stacking

Within the field of drug development, Quantitative Structure-Activity Relationships (QSARs) are used extensively in a similar manner to the multidimensional LFERs described above. QSAR builds on the work of Hammett and others on linear free energy relationships, expanding the concept to multiple descriptors in order to predict drug potencies. Over time, as machine learning and statistical methods have been developed and gained popularity, these relationships have become incredibly complex, and difficult to interpret. In general, they cannot be easily used to gain any understanding of drug binding. While explicitly computing properties that occur at the cellular level is computationally intractable, quantum mechanical based computational tools are well suited to quantifying specific intermolecular interactions, which can provide insight into the driving factors behind experimentally observed drug binding trends. I have developed multi-parameter LFERs to predict the strength of non-covalent interactions important to drug binding with the aim of *retaining the simplicity necessary to provide chemical insights into these interactions*. These relationships can then be used in fragment-based drug design and to improve scoring functions for docking methods while also providing us with improved chemical understanding of the underlying non-covalent interactions.

Stacking interactions, though common in drug binding sites, are not well understood and are thus a good candidate for this type of analysis. Past work found correlations between heterocycle dipole moments and interaction energies for stacking

intrdactions with both benzene[30] and the protein amide backbone.[18] However, these correlations were weak, and when a larger and more diverse set of heterocycles was used, dipole moment was no longer sufficient to describe the observed trends.[31] This highlighted the need for both fits that can consider multiple heterocycle features and improved descriptors of heterocycle electrostatic features beyond simply the dipole moment.

Past work using LFERs in the understanding of heterocycle interactions relevant to drug binding have focused on the use of dipole moment to capture electrostatic effects.[18, 30-31] However, it is known that molecular dipoles only accurately represent intermolecular interactions at large distances, and not the short distances relevant understanding interactions in a protein binding pocket. Further, there are a large number of heterocycles with zero dipole moment due to their symmetry, but with very different interaction strengths due to their distinct substitution patterns that cannot be captured using dipole moments alone as descriptors (i.e. the difference between benzene and 1,3,5-triazine). Thus, I developed new, physically motivated molecular descriptors that can be used to describe the electrostatics of heterocycle drug fragments at the distances relevant to drug-protein interactions. Additionally, I have applied these new molecular descriptors to the development of LFERs that predict and explain many of the different types of interactions that heterocyclic drug fragments can participate in, which have unique trends in their tunability and geometric preferences. In particular, I focused on interactions that have been observed in drug binding sites and are expected to have a significant electrostatic component.

My first study in this area concerned the development of a multidimensional LFER for the prediction of interaction energies of amide stacking with heterocycles.[2-3] This "non-traditional" stacking interaction can be quite strong, and is expected to play a role in protein structure and stability,[4-6] urea-induced protein denaturation,[7] and protein-ligand binding;[8-10] however the interaction itself is not well understood. Optimization of amide stacking interactions has been used to improve inhibition of several protein targets including serine protease Factor Xa,[8] endothiapepsin,[11-12] protein kinase A (PKA),[13] and the cysteine protease autophagin-1.[14] It is also an appealing drug design target as the amide backbone is generally less flexible than the side chains and the large number of intra-protein amide hydrogen bonds leave the π-face as a more accessible target. [15-17]

My second study focused on the interaction between planar salt-bridges and heterocycles. These interactions have been observed in protein-ligand interactions[20] and are also common at protein-protein interaction interfaces.[21-23] Experimental studies of synthetic models systems have shown that the presence of aromatic groups promotes salt-bridge formation.[24-26] Similar to amide stacking, salt-bridge stacking interactions are expected to be electrostatically controlled. However, rather than the large molecular dipole observed in an amide, in the case of the salt bridge there are separated positive and negative charges. The acetate-guanidinium ion pair was used as a model of an Asp:Arg salt bridge, and due to the larger size of the ion pair, the heterocycle set was expanded to include 27 bicyclic and three tricyclic heterocycles.

Thirdly, I have studied stacking interactions between a subset of the monocycles and bicycles with the aromatic amino acid side chains phenylalanine, tyrosine and tryptophan. These amino acids are modeled by toluene, *p*-methylphenol, and 9-

methylindole, respectively. While arene stacking interactions are often a key component of drug binding, and small changes in a drug heterocycle that interacts with aromatic residues have been shown to dramatically impact binding,[27] the reasons for these changes are not well understood. There are not currently any tools available to help medicinal chemists make judicious choices regarding the incorporation of different heterocycles based on their stacking ability. The development of a quantitatively predictive LFER, along with interpretation of that LFER into clear qualitative design guidelines, will help to guide this decision making.

The LFERs created to describe these three interaction types are all based on our newly developed electrostatic potential based molecular descriptors, in combination with terms to capture dispersion. While these fits offer useful predictions at a fraction of the computational cost of full dimer geometry optimization, the quantum chemical evaluation of these descriptors may still be prohibitive for truly high throughput prediction of heterocycle performance. Two of these descriptors, ESP_{mean} and ESP_{max}, have a direct relationship with the amount of substitution on the heterocycle and the topological arrangement of these heteroatoms. Thus, they can be evaluated directly from a description of atom connectivity such as a SMILES string. Incorporation of these connectivity-derived molecular descriptors into the previously developed LFER for arene stacking leads to predictions for interaction energies of similar quality to those based on the QM computed descriptors.

Motivation for Studying a Conjugate Addition

The use of multivariate regression in asymmetric catalyst design has been recently popularized by Sigman and coworkers and successfully applied to a variety of

reactions.[32, 34-44] They have further shown that despite the increased complexity of their parametrized function, resulting from the inclusion of a larger number of parameters, these fits can still be used to gain mechanistic understanding of the stereo-determining steps.[34, 38-39] Such understanding comes from analysis of which catalyst descriptors of those considered are shown to be important via the fitting and parameterization process. However, development of these LFERs requires access to a large number of experimental results for catalysts that span a large range of selectivities. This may restrict this type of analysis to reaction types where catalyst synthesis is straightforward, modular, and amenable to a variety of substitutions. This provides an opportunity for further application of computational methods, in particular density function theory (DFT), which has been shown to be useful in understanding asymmetric reactions and guiding catalyst design.[45-52] Since computational results can provide accurate rates and selectivities, the use of computational data in multivariate LFERs in cases where experimental data is not available is appealing. Further, development of the program AARON within our group has automated the process of transition state optimization, overcoming problems with the time intensive nature of that process and allowing for a quantum mechanical virtual screening. This leads to the creation of sufficient computational data for the development of an LFER.

This methodology was applied toward the design of an improved catalyst for the enantioselective synthesis of α-chiral heterocycles via the conjugate addition of vinyl boronic acids to heterocycle appended enones. The reaction, introduced by May et al.[54] in 2012, is catalyzed by a 3,3'-substituted BINOL, and shows significant dependence of both rate and selectivity on the identity of the 3,3' substituent. Virtual screening for

identification of an improved catalyst was combined with the development of an LFER that could identify key features of the substituent that contributed to the overall catalyst performance.

CHAPTER 2

STACKING INTERACTIONS OF HETEROCYCLIC DRUG FRAGMENTS WITH PROTEIN AMIDE BACKBONES [a]

[a] Bootsma, A. N and S. E. Wheeler. *ChemMedChem* **13**,835 (2018).
Reprinted here with permission of the publisher

Abstract

Stacking interactions can be important enthalpic contributors to drug binding. Among the less well-studied stacking interactions are those occurring between an arene and the π-face of an amide group. Given the ubiquity of heterocycles in drugs, combined with the abundance of amides in the protein backbone, optimizing these non-covalent interactions can provide a potential route to enhanced drug binding. Previously, Diederich et al. (ChemMedChem 2013, 8, 397-404) studied stacked dimers of a model amide with a set of 18 heterocycles, showing that computed interaction energies correlate with the dipole moments of the heterocycles and providing guidelines for the optimization of these interactions. We have considered stacked dimers of the same model amide with a larger set of 28 heterocycles common in pharmaceuticals using more robust ab initio methods. While the overall trends in these new data corroborate many of the results of Diederich et al., these data provide a more refined view of the nature of amide stacking interactions. We present a robust scoring function for amide stacking interaction energies based on the molecular dipole moment and strength of the electric field above the arene.

Introduction

Stacking interactions, which are typically defined as occurring between two aromatic rings,[55-56] play critical roles in fields ranging from structure-based drug design and molecular recognition to the development of organic electronic materials and asymmetric catalysts.[57-64] However, there are a growing number of "non-traditional" stacking interactions that have proved important in protein-ligand interactions.[60] Among these are amide stacking interactions,[2-3] which are the attractive interactions between

arenes and the π-system of an amide. Amide stacking interactions are stronger than many other stacking interactions, and can be competitive with hydrogen bonds;[65-66] they can also play a role in protein structure and stability,[4-6] and similar interactions have been implicated in urea-induced denaturation of proteins.[7] Unlike stacking interactions of arenes,[30-31, 55-56, 67-76] amide stacking is still relatively poorly understood.

Stacking interactions of amides with heterocycles are expected to play an important role in drug binding due to the abundance of backbone amide groups in proteins and the ubiquity of heterocycles in drugs.[31, 77-79] Furthermore, since the protein backbone is generally less flexible than the amino acid side chains, it should be simpler to target amide stacking interactions through structure-based drug design.[18] Additionally, the large number of intra-protein hydrogen bonds involving backbone amides means that the π-face of these functional groups is typically more accessible for interactions with ligands.[15-17]

Amide stacking interactions have been implicated in the binding of a number of small-molecule inhibitors.[8-10] For instance, in 2005, Perzborn *et al*.[8] demonstrated significantly increased inhibition of serine protease Factor Xa (FXa) by introducing an amide into an oxazolidinone based inhibitor such that it could stack between Phe and Tyr side chains in the binding pocket. In 2012, Diederich *et al*.[9] also took advantage of amide stacking in the inhibition of FXa, showing that optimizing the orientation of an oxazole fragment in a small-molecule inhibitor, relative to a backbone amide in the binding site, improved binding by an order of magnitude. A recent computational study by Sherrill *et al*.[10] comparing two FXa targeting ligands[80] found that differences in interaction energy stem largely from differences in amide stacking.

In 2013, Diederich *et al*.[18] presented computational data and results from an analysis of the PDB regarding stacking interactions between a model amide (*N*-methylacetamide, NMAC) and 18 heterocycles common in pharmaceuticals. Ultimately, they provided three guidelines for choosing heterocyclic drug fragments that yield strong stacking interactions with amides:

1. orienting the fragment dipole vector such that it is aligned in an antiparallel fashion with the dipole of the interacting amide group
2. increasing the dipole moment of the heterocycle
3. decreasing the π-electron density of the heterocycle (*i.e*. through the introduction of nitrogens).[81]

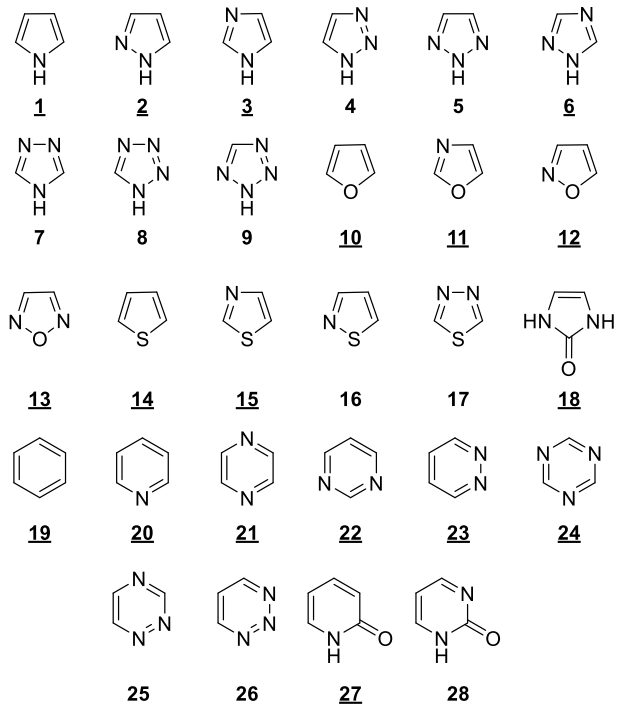
Optimizing amide stacking has also been used to improve inhibition of a number of targets, including the aspartic protease endothiapepsin,[11-12] protein kinase A (PKA),[13] and the cysteine protease autophagin-1.[14] Most recently, Diederich *et al*.[29, 82-83] tried to design an inhibitor for the cysteine protease cathepsin L by maximizing amide stacking interactions. Variations of a heterocyclic core led to K_i values spanning two orders of magnitude (from 1450 nM for quinolone to 4 nM for benzothiophene). However, they found no clear relationship between the hetereocycle dipole orientations and measured K_i values. This was attributed to the presence of two amides in the binding pocket with antiparallel dipole moments in close proximity to the heterocycle, which is expected to buck any trends extracted from simpler model amide stacking interactions.

The first two of Diederich's guidelines[18] was based on a correlation between computed interaction energies for model stacked dimers and the molecular dipole moments of the heterocycles ($r^2 = 0.84$). However, this correlation excluded heterocycles

with zero dipole (*i.e.* benzene, pyrazine, and 1,3,5-triazene). Moreover, it applied only to sandwich and parallel-displaced dimers separately;[84] once all heterocycles are included, and the global minimum energy structure reported for each heterocycle considered (regardless of whether it is in a sandwich or parallel-displaced configuration), the correlation of the heterocycle dipole moments with the interaction energies[18] is considerably weaker ($r^2 = 0.50$; see Appendix A, Figure A-1). This suggests that simply using the heterocycle dipole moment as a predictor of amide stacking interactions will be potentially problematic, particularly in cases in which the geometry of the amide stacking interaction (*i.e.* sandwich vs. parallel displaced) is not already precisely known.

Subsequent work by Huber and co-workers[30] on stacking interactions between a set of 10 monocyclic heterocycles and benzene showed a similar correlation of computed stacking energies with hetereocycle dipole moments ($r^2 = 0.78$). However, An, *et al.*[31] showed that this correlation is diminished when looking at stacking of 9-methyladenine with a set of 28 monocyclic heterocycles ($r^2 = 0.73$). Furthermore, for stacking of 9-methyladenine with larger fused heterocycles this correlation all but disappears ($r^2 = 0.29$).

Here, we provide a more robust *ab initio* computational study of stacking interactions between NMAC and common heterocyclic drug fragments, employing both higher levels of *ab initio* theory and a larger and more diverse set of heterocycles than considered by Diederich *et al.*[18] (see Scheme 2.1). Ultimately, we present a predictive model of amide-heterocycle interaction energies based on simple molecular descriptors that can account for all heterocycles and preferred stacked geometries, as well as qualifications for the design guidelines put forward by Diederich *et al.*[18]



Scheme 2.1. Heterocycles included in this study. Underlined numbers denote heterocycles that were also considered by Diederich *et al.*[18]

Results and Discussion

We considered the set of 28 heterocycles in Scheme 2.1, which includes the 18 studied by Diederich *et al.*[18] augmented with other common monocyclic heterocycles found in pharmaceuticals as well as their tautomers and regioisomers.[31, 77-79] In order to ensure all stacked energy minima were identified, we considered 12 distinct starting geometries for each heterocycle (see Theoretical Methods for more details). Following Diederich *et al.*,[18] constrained geometry optimizations were performed in which the heavy atoms of the amide and heterocycle were restricted to remain in parallel planes while all other geometrical parameters were optimized. These optimizations were performed at the ωB97X-D/def2-TZVP level of theory,[85] which should provide robust optimized geometries.[86] The resulting stacked dimers were characterized by three

intermolecular coordinates (see Figure 2.1). The interaction energies of these complexes were then computed at the DLPNO-CCSD(T)/cc-pVQZ[87-90] level of theory (see Table 2.1), which should provide benchmark-quality data (See Appendix A, Table A-1 for more information and comparisons with other, even higher levels of theory). Excellent agreement of these benchmark interaction energies with the ωB97X-D/def2-TZVP data (RMSE = 0.4 kcal mol^{-1} across all 116 local stacked energy minima; see Appendix A, Figure A-2) supports the use of the latter level of theory for the geometry optimizations.

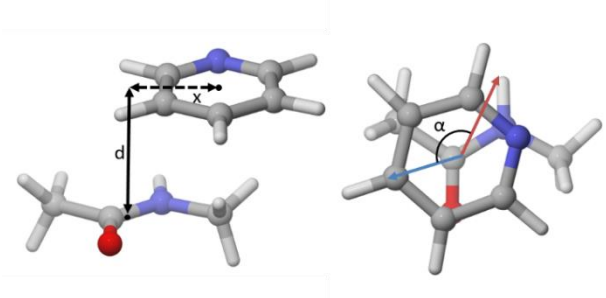


Figure 2.1. Intermolecular coordinates (d, x, and α) characterizing stacked amide-heterocycle dimers, shown for model dimers of pyridine and n-methylacetamide (NMAC). The blue and red vectors represent the dipole moments of pyridine and NMAC, respectively.

There are qualitative differences in the structures and computed interaction energies between the present work and those reported by Diederich *et al*.[18] based on LMP2/cc-pVDZ optimized geometries. In particular, whereas they found that the optimized dimers naturally segregated into sandwich type dimers (x < 0.7 Å) and parallel displaced structures (x > 1.1 Å), we find a nearly continuous distribution of parallel-displaced distances centred at 1.0 Å (see Appendix A, Figure A-3 for more details). In other words, we find no natural means of separating sandwich from parallel displaced geometries among the minimum energy configurations for each heterocycle. Notably, this is not an artefact of the particular DFT functional employed here, but is reproduced by a

number of reliable DFT functionals (see Appendix A, Table A-2). We also find a much smaller range of interplanar distances (3.17 to 3.41 Å) compared to Diederich *et al*. (3.20 to 3.81 Å) and much stronger overall interaction energies (average interaction energy of -5.5 kcal mol^{-1} in our data compared to -2.8 kcal mol^{-1} from Diederich, for the 18 heterocycles common to both studies).

Despite these differences, our more robust data corroborate the finding of Diederich *et al*.[18] that stacked dimers of amides with heterocycles tend to adopt orientations with the molecular dipoles in an antiparallel orientation (see Figure 2.2A). However, the distribution of α-values over this larger set of heterocycles is broad, with over a third of the heterocycles exhibiting α-values less than 150°. That is, while there is a general tendency of amides to stack in an antiparallel arrangement, the trend is weak. This can be understood by plotting the optimized α-values against the heterocycle dipole moments (see Figure 2.2B). As expected, while there is a strong tendency for the heterocycles with large dipole moments to align in an antiparallel fashion, this tendency quickly dies off with decreasing dipole. The result is that for heterocycles with smaller dipoles, α-values that deviate significantly from 180° are common (for heterocycles with zero dipole moment this angle is not defined).

The broad distribution of α-values can also be understood by explicitly considering the dependence of the interaction energy on α for selected systems. Interaction energies as a function of α are plotted in Figure 2.2C for pyridazine and pyridine. In the latter case, we considered two starting geometries for these scans, corresponding to slightly different lateral displacements (x). For systems such as pyridazine with a large dipole moment (μ = 4.24 D), there is a relatively steep energetic

penalty for deviations from α = 180°, which reaches a maximum of 5.4 kcal mol^{-1} for α = 0°. However, for the many heterocycles with more modest dipole moments (*e.g.* pyridine, μ = 2.25 D), the angular dependence of the interaction energy is weak, with a very broad energy minimum surrounding α = 180°. As seen in Figure 2.2C, the shape of this broad part of the potential varies slightly depending on the lateral displacement, leading to local energy minima with widely varying α-values. As discussed more thoroughly below, for these heterocycles with modest dipole moments, direct interactions between the heterocycle and amide can easily overwhelm the overall tendency of the molecular dipoles to align in an antiparallel orientation.

Looking at the interaction energies in Table 2.1 in more detail, a number of examples emerge that run counter to the design guidelines of Diederich *et al.*,[18] suggesting the need for some refinement of our understanding of these interactions. For example, although 1,2,3-triazole has a much larger dipole moment than 1,2,4-triazole (see Figure 2.3) their interaction energies with amides are essentially identical. Two factors contribute to this. First, the optimized stacked dimer of NMAC with 1,2,4-triazole exhibits two N···H_3C interactions, while only one such interaction is observed in the 1,2,3-triazole case. Second, 1,2,4-triazole can access both interactions while maintaining an antiparallel dipole orientation (α = 171°), but in order for 1,2,3-triazole to engage in an N···H_3C interaction it must be rotated away from antiparallel geometry (α = 154°).

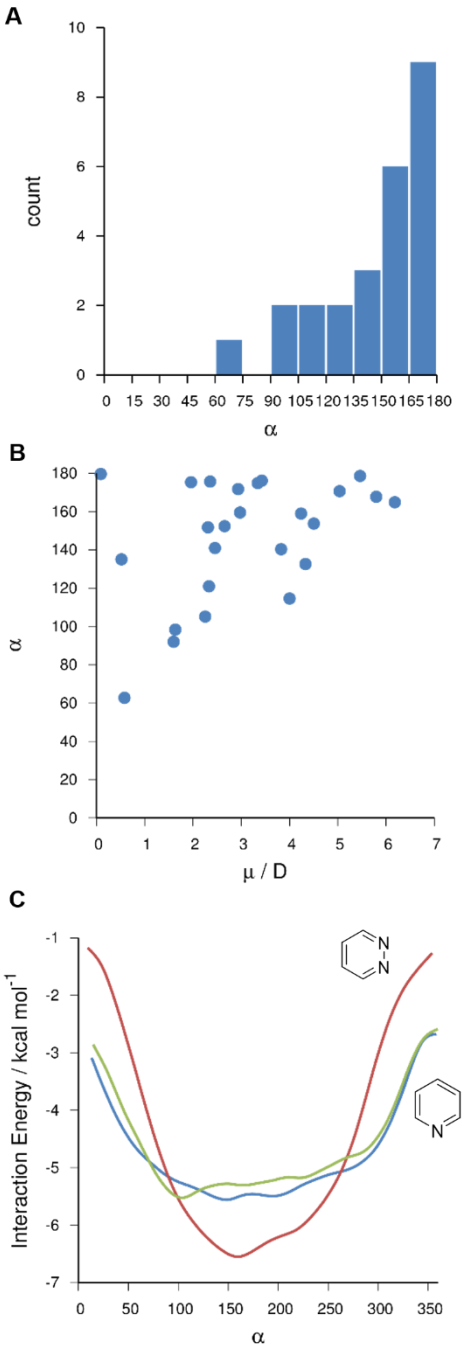


Figure 2.2. A. Histogram of computed α values among minimum energy dimer structures. B. Scatter plot of the angle α vs the heterocycle dipole moment, μ, for the minimum energy dimers. C. ωB97X-D/def2-TZVP Interaction energies (kcal mol^{-1}) as a function of α for NMAC stacked with pyridazine (red) and pyridine (blue and green) starting from two different local energy minima.

The isoelectronic series of heterocycles going from benzene to 1,3,5-triazine provide even greater insight into the nature of amide stacking interactions, underscoring the importance of local interactions. First, one sees a 0.4 kcal mol^{-1} enhancement in interaction energy upon addition of a nitrogen to go from benzene to pyridine. This is consistent with the guidelines of Diederich *et al.*,[18] since pyridine has a dipole while benzene does not. However, in this global minimum energy structure, the dipole moment of pyridine is nearly perpendicular to that of NMAC ($\alpha = 105°$)! As such, the enhanced amide stacking of pyridine cannot be attributed to the interaction of the molecular dipole moments. Instead, it can be explained by the presence of a favourable N···H_3C interaction, which is obviously absent in the case of benzene. The addition of a second nitrogen leads to similar improvement as was seen with the first, regardless of whether one considers pyrimidine ($\mu = 2.35$ D) or pyrazine ($\mu = 0$ D), as both exhibit an additional N···H_3C interaction. In the case of pyrimidine, one sees an antiparallel alignment of the heterocycle dipole with that of NMAC ($\alpha = 176°$). However, this is primarily a consequence of the local N···H_3C interactions. Finally, one sees no difference in interaction energy for 1,3,5-triazine compared with pyrimidine and pyrazine. This occurs because there is no available CH_3 group with which the third nitrogen in 1,3,5-triazine can interact. The differences in geometry seen among the global minimum energy structures for these six-membered heterocycles can also be explained in terms of direct interactions. Pyridine and pyrazine maintain the geometry of the benzene-amide dimer, since this is compatible with forming two N···H_3C interactions, whereas maintaining these two interactions requires a lateral displacement of pyrimidine and 1,3,5-triazine, relative to the amide.

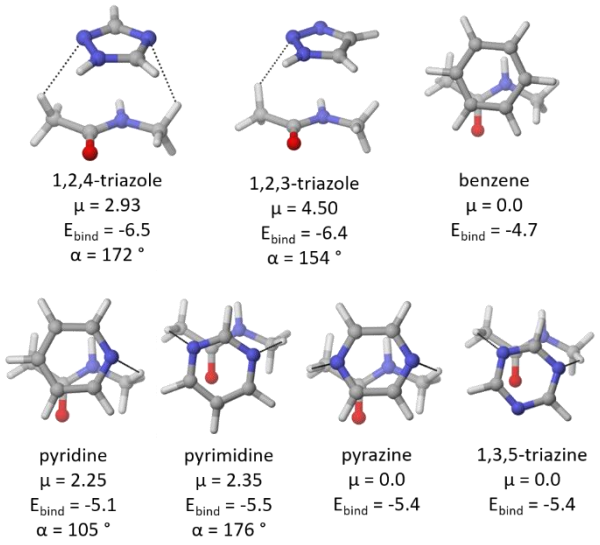


Figure 2.3. Global minimum energy dimers for selected heterocycles (dipole moments in D, E_{bind} in kcal mol^{-1}).

Despite the variations in interaction energies arising from direct interactions between the heterocycle and amide, and the relative unimportance of molecular dipole-dipole interactions, we sought to develop a simple scoring function that provides reliable predictions of amide-heterocycle interaction energies regardless of the geometry of the resulting complex. First, we note that the more robust *ab initio* interaction energies provided here, including the global minimum structure for each heterocycle as well as heterocycles with zero dipole, are more strongly correlated with the heterocycle dipole moments than reported by Diederich *et al.*[91] ($r^2 = 0.71$, see Figure 2.4A compared to Figure A-1). That is, these new interaction energies appear to provide stronger support for Diederich's main conclusion regarding the importance of molecular dipole moments, similar to what has been previously reported for heterocycles stacking with benzene[30] and 9-methyladenine. [31] However, as discussed above, it appears that the heterocycle dipole serves primarily as an indirect indicator of the presence of heteroatoms capable of engaging in direct interactions with the amide.

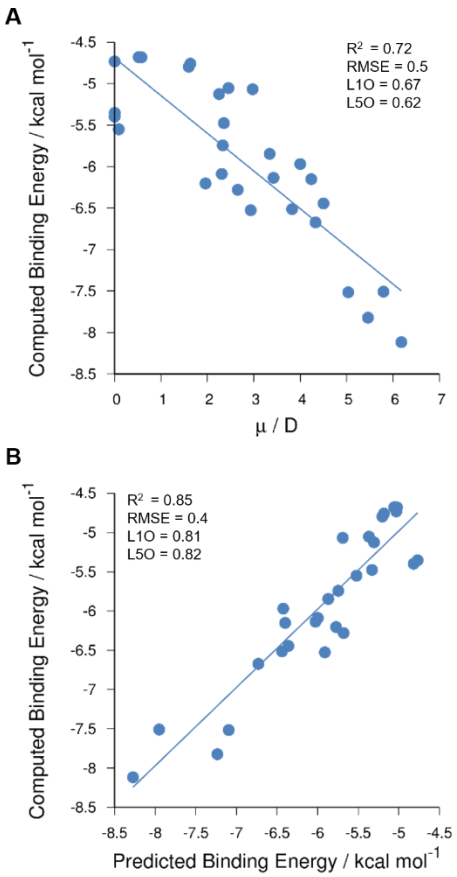


Figure 4. Computed interaction energies vs (A) Dipole and (B) interaction energy predicted using equation (2.1).

In order to identify a more robust predictor of amide interaction energies, we performed multi-parameter linear regressions based on a number of heterocycle descriptors. SAPT0 energy decompositions of the interactions for all local energy minima reveal that the trend in interaction energies among heterocycles is primarily a function of electrostatic effects. This is consistent with the recent F-SAPT results from Sherrill and coworkers.[10] Thus, parameters that capture electrostatic properties of the heterocycles were investigated most thoroughly, including parameters that describe both local and global electrostatic character. Ultimately, the following two-parameter fit provided the simplest but most reliable predictor,

$$E_{bind} = -0.067 * \mu^2 - 1.06 * \left|\vec{E}\right|_{max} - 4.35. \quad (2.1)$$

In equation (2.1), the dipole moment is squared to account for the increased importance of larger dipole moments in determining the interaction energy while $|\vec{E}|_{max}$ is the maximum electric field magnitude in the plane 3.25 Å above the heterocycle. The latter parameter is able to capture local variations in the electrostatic potential more effectively than the molecular dipole moment, allowing for predictions of amide stacking interactions even for highly polarized systems with zero net dipole. Values for these parameters, and the resulting energy predictions are listed in Table 2.1.

DLPNO-CCSD(T) interaction energies are plotted against those predicted using equation (1) in Figure 2.4B. The data are strongly correlated ($r^2 = 0.85$; $q^2 = 0.81$), with an RMSE of 0.4 kcal mol^{-1}. The accuracy of this simple fit confirms that an electrostatic explanation is adequate to describe the observed trends, while simultaneously showing the need for more local information than that provided by the molecular dipole moment alone. Moreover, cross validation via leave-one-out and leave-5-out analyses reveals that the two-parameter fit in Figure 2.4B is more robust than that in Figure 2.4A. Apparently, inclusion of descriptors of local electrostatic characteristics (via the electric field) is sufficient to capture the direct interactions discussed above (Figure 2.3). Ultimately, the strong correlation between the predictions from equation (1) and the reference *ab initio* data support a physical picture of amide stacking in which local electrostatic effects are of paramount importance. We have previously reported similar findings for stacking interactions involving heterocycles.[31, 74, 92] This simple predictor should provide a more reliable scoring function for choosing heterocyclic fragments expected to maximize amide stacking interactions in drug binding sites.

Conclusions

Interaction energies for stacked dimers of a model amide (NMAC) with 28 common heterocyclic drug fragments were determined at the DLPNO-CCSD(T)/cc-pVQZ//ωB97X-D/def2-TZVP level of theory. Interaction energies for these dimers range from -4.7 to -8.1 kcal mol^{-1}. These interaction energies are substantial, providing a potentially powerful means of enhancing protein-ligand binding enthalpies. Of course, these interactions in the bound state will be in competition with the cost if ligand desolvation.

Electrostatic effects control energetic differences amongst heterocycles through a combination of local dipole moments and direct interactions, both of which impact the preferred stacking orientation. In general, the heterocycle is oriented such that the heterocycle dipole moment and amide dipole moment are close to antiparallel, but local interactions such as N···H_3C modulate this preference. This effect may be complicated in real binding pockets, as there are many, potentially stronger, competing effects of adding nitrogen atoms to ligands.[93] Ultimately, we have shown that incorporating molecular descriptors based on computed electric fields into the model to account for local effects provides a more sound predictor of the strength of amide stacking interactions than the purely dipole-based model presented by Diederich *et al*.[18] Furthermore, this two component model is able to explain trends in interaction energies of amides with arenes that have no net dipole moment.

Table 2.1. Dipole moments (μ), the maximum of the electric field in the plane above the heterocycle ($\left|\vec{E}\right|_{max}$, in kcal mol^{-1} D^{-1}), and intermolecular coordinates for the global minimum energy stacked dimers of NMAC with heterocycles **1**-**28**, as well as the computed and predicted interaction energies (E_{bind}) from Eq. (1).

						E_{bind} (kcal mol^{-1})	
Het.	μ (D)	$\left\|\vec{E}\right\|_{max}$	α (°)	d (Å)	x (Å)	Comp.	Pred.
1	1.96	0.99	175	3.17	1.51	-6.2	-5.8
2	2.33	0.62	121	3.24	1.28	-5.7	-5.7
3	3.83	1.10	140	3.20	1.09	-6.5	-6.4
4	4.50	0.93	154	3.27	0.84	-6.4	-6.4
5	0.09	1.29	180	3.21	1.26	-5.5	-5.5
6	2.93	0.73	172	3.24	0.84	-6.5	-5.9
7	5.79	0.81	168	3.21	0.71	-7.5	-8.0
8	5.46	0.39	179	3.20	0.56	-7.8	-7.2
9	2.31	0.64	152	3.23	0.95	-6.1	-6.0
10	0.58	0.95	63	3.23	1.37	-4.7	-5.0
11	1.60	0.61	92	3.20	1.26	-4.8	-5.2
12	2.97	1.05	159	3.28	0.84	-5.1	-5.7
13	3.34	0.58	175	3.27	0.57	-5.8	-5.9
14	0.52	0.70	135	3.33	1.57	-4.7	-5.1
15	1.63	0.64	98	3.41	0.85	-4.8	-5.2
16	2.45	0.44	141	3.34	1.34	-5.1	-5.4
17	3.42	0.97	176	3.19	1.48	-6.1	-6.0
18	4.00	0.80	115	3.26	1.14	-6.0	-6.4
19	0.00	0.58	--	3.34	1.36	-4.7	-5.0
20	2.25	1.06	105	3.33	1.08	-5.1	-5.3
21	0.00	1.30	--	3.31	0.96	-5.4	-4.8
22	2.35	0.57	176	3.32	0.88	-5.5	-5.3
23	4.24	1.10	159	3.31	0.56	-6.2	-6.4
24	0.00	0.84	--	3.20	1.50	-5.4	-4.8
25	2.65	1.21	152	3.30	0.76	-6.3	-5.7
26	5.04	0.84	171	3.21	1.09	-7.5	-7.1
27	4.33	0.61	133	3.32	0.81	-6.7	-6.7
28	6.18	0.64	165	3.25	0.28	-8.1	-8.3

Overall, these data suggest a number of caveats for the guidelines put forth by Diederich *et al*.[18] First, the preference for an antiparallel orientation of dipole moments in amide stacking interactions is only significant for heterocycles with very large dipole moments. Second, for most systems the effect of molecular dipole is easily overwhelmed by specific local interactions, such as N···H_3C interactions, which contribute 0.3 to 0.5

kcal mol^{-1} each. Finally, increasing the number of nitrogens in a given heterocycle only improves the amide interaction energy if the additional nitrogen(s) can engage in a local interaction with the amide. Remarkably, these sundry effects can all be captured with a simple two-parameter model (Eq 2.1).

Local interactions have been shown to be key contributors to stacking interactions with a variety of aromatic systems.[31, 74, 92] In the case of amide stacking, these effects are captured adequately by the maximum electric field in the plane above the heterocycle; inclusion of this second heterocycle descriptor improves the correlation compared to the dipole moment alone, increasing r^2 from 0.72 to 0.85. Although both Huber *et al*.[30] and An *et al*.[31] demonstrated that molecular dipole moments alone can predict the arene stacking interactions with monocyclic heterocycles, this correlation breaks down when considering bicyclic heterocycles.[31] Whether the incorporation of the electric-field strength of the heterocycles will be sufficient to rescue this correlation will be addressed in future work.

We hope that these refined design guidelines, combined with the simple predictive model in equation (2.1), will prove useful in choosing heterocyclic drug fragments that maximize amide stacking interactions.

Computational Methods

Starting with the 18 heterocycles originally studied by Diederich and co-workers,[18] we added ten additional structures chosen to ensure all regioisomers and tautomers were included. The aim of expanding this set of heterocycles was to provide a clearer view of broad trends and more reliable cross-validation of predictive models. Geometries were optimized using ωB97X-D/def2-TZVP[85] starting at 12 distinct starting

geometries. To obtain these starting geometries the heterocycle was placed 3.5 Å above the plane of the amide either with the centres of mass aligned, or with a 1.5 Å shift. The heterocycle was then rotated around the axis through the centre of mass of the amide in 60° increments to generate 6 starting conformations at each shift. Optimizations were performed with a constraint to maintain planarity of the heavy atoms of the amide and the heterocycle, and to keep those two planes parallel to one another. After optimisation structures were screened for uniqueness, with an RMSD cut-off of 0.4 Å. This resulted in 2 to 7 unique dimers for each heterocycle. DLPNO-CCSD(T)/cc-pVQZ[87-90] single point energies were computed for all local minima (116 across all heterocycles). The interaction energy was calculated as the difference in energy between the optimized dimer and the optimized isolate monomers (optimised at the ωB97X-D/def2-TZVP level of theory). Counterpoise corrections were not used. All DFT computations were performed using Gaussian 09,[94] while DLPNO-CCSD(T) energies were performed in ORCA 4.0.1[95].

The maximum electric field in the plane above each heterocycle, $\left|\vec{E}\right|_{max}$, was computed by considering a 10 x 10 Å grid of points 0.1 Å apart, in the plane 3.25 Å above the plane of the heterocycle. This distance (3.25 Å) corresponds to the average of the d-values for the dimers considered; however, we note that the trends shown are very weakly dependent on this distance. Dipole moments were determined computationally at the ωB97X-D/def2-TZVP[85] level of theory. Equation 2.1 was determined by least squares linear regression. These fits were cross validated using scikit-learn version 17.1[91] by performing leave-one-out (L1O) and leave-five-out (L5O) analyses.

CHAPTER 3

TUNING STACKING INTERACTIONS BETWEEN ASP-ARG SALT BRIDGES AND HETEROCYCLIC DRUG FRAGMENTS [b]

[b] Reprinted with permission from Bootsma, A. N and S. E. Wheeler. *J. Chem. Inf. Model.* **59**, 149 (2019). Copyright (2019) the American Chemic Society

Abstract

Stacking interactions can play an integral role in the strength and selectivity of protein-drug binding and are of particular interest given the ubiquity and variety of heterocyclic fragments in drugs. In addition to traditional stacking interactions between aromatic rings, stacking interactions involving heterocyclic drug fragments and protein salt-bridges have also been observed. These 'salt-bridge stacking interactions' can be quite strong, but are not well understood. We studied stacked dimers of the acetate···guanidinium ion pair with a diverse set of 63 heterocycles using robust *ab initio* methods. Computed interaction energies span more than 10 kcal mol^{-1}, indicating the sensitivity of these salt-bridge stacking interactions to heterocycle features. Trends in both the strength and preferred geometry of these interactions can be understood through analyses of the electrostatic potentials and electric fields above the heterocycles. We have developed new heterocycle descriptors that quantify these effects and used them to create robust predictors of the strength of salt-bridge stacking interactions both in the gas phase and a protein-like dielectric environment. These predictive tools, combined with a set of qualitative guidelines, should facilitate the choice of heterocycles that maximize salt-bridge stacking interactions in drug binding sites.

Introduction

Non-covalent interactions involving aromatic rings play a significant role in fields ranging from structure-based drug design and molecular recognition to asymmetric catalysis and the solid-state packing of organic electronic materials.[57, 59-61, 96-97] These interactions include the well-known stacking interactions that occur between aromatic rings[55-56] as well as many "non-traditional" stacking interactions.[60] Stacking interactions

are important for the binding of ligands by proteins and provide a means by which ligand binding enthalpies can be tuned over a considerable range.

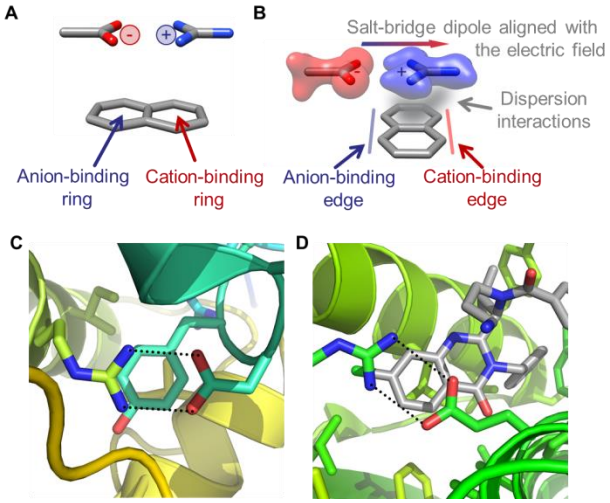


Figure 3.1. A) A simple model of salt-bridge stacking in which different rings of a polycyclic arene bind the cation and anion; B) A more refined model of salt-bridge stacking over a bicyclic heterocycle based on observed structural trends; C) Asp-Arg salt-bridge stacking over a Tyr side chain in the feruloyl esterase module of xylanase 10B (PDB ID: 1GKK); D) Glu-Arg salt-bridge stacking over the quinazolinone core of the KSP inhibitor ispinesib (PDB ID: 4A5Y).

Among the less well-studied stacking interactions, particularly in the context of structure-based drug design, are those occurring between salt-bridges (*e.g.* Asp-Arg dyads) and polycyclic arenes (see Figure 3.1). These interactions have been studied in small synthetic model systems[24-26] and discussed in a number of contexts under different names. Mukhopadhyay and coworkers[98] introduced the term 'salt-bridge-π interaction' to describe stacking interactions between a planar salt-bridge and an aromatic ring, and then demonstrated these interactions in inorganic-organic hybrid crystals. Matile and co-workers[99-101] subsequently reported that polycyclic arenes can simultaneously bind an anion and cation in what they termed 'ion-pair-π interactions,' showing that these interactions are most favorable when there is an anti-parallel alignment of the salt-bridge and heterocycle dipole moments. Herrera *et al.*[102] demonstrated analogous intermolecular

interactions in an organocatalytic reaction, introducing the term 'push-pull π+/π- interaction.' Here, we use the term 'salt-bridge stacking interaction,' which can be considered a specific type of Matile's 'ion-pair-π interaction',[99-101] and focus on understanding the means by which these interactions can be harnessed in the context of the protein-ligand interactions.

The stacking of salt-bridges with arenes has been observed in a number of biological contexts.[20-23, 103-105] For instance, cation-anion-aromatic triads are commonly observed at protein-protein interfaces.[23] This can be seen in the interaction of an importin β transport receptor with GTP-binding protein Ran[21] and in the binding of HIV-1 Nef with the SH3 domain of Src kinases.[22] The stacking of an Asp-Arg salt-bridge with a Tyr side chain is also apparent in the feruloyl esterase module of xylanase 10B from *C. thermocellum* (see Figure 3.1C).[104] Salt-bridge stacking interactions have also been noted in protein-ligand binding. For instance, the quinazolinone core of the kinesin spindle protein inhibitor ispinesib (SB-715992) stacks with a Glu-Arg salt-bridge in mitotic kinesin Eg5 (see Figure 3.1D).[20] Despite the ubiquity of these interactions, and their demonstrated utility, the nature of salt-bridge stacking interactions and the means by which these interactions can be tuned have not been well-studied.

Previous work has shown the utility of univariate and multivariate linear regressions in describing the strength of stacking interactions involving heterocyclic drug fragments and in understanding their physical underpinnings.[18, 30-31, 106] For instance, in 2013, Diederich *et al.*[18] showed that heterocycle dipole moments provide a qualitative descriptor of the strength of stacking interactions between heteroaromatic drug fragments and protein amide backbones. We subsequently developed[106] a two-parameter predictive

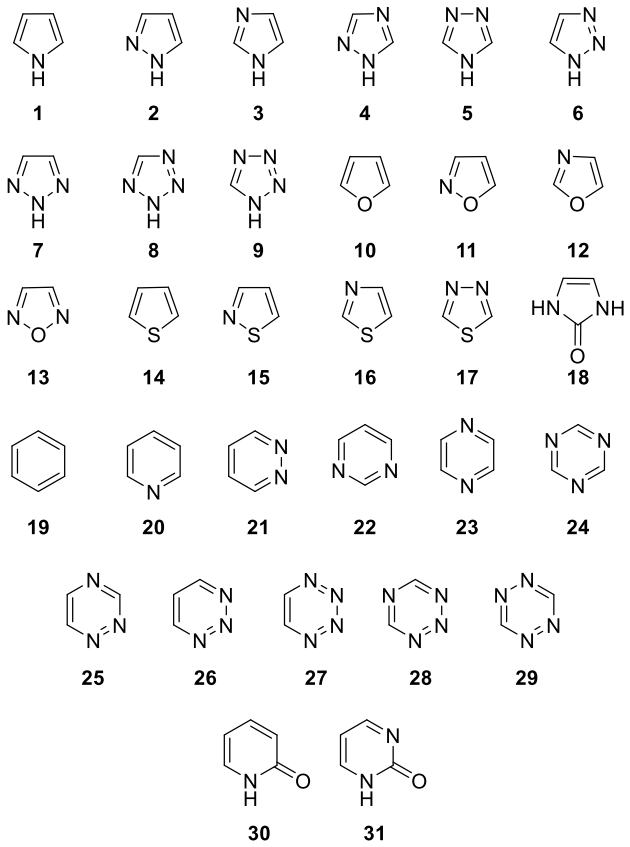
model of these 'amide stacking interactions' based on the heterocycle dipole moment and a new heterocycle descriptor based on the electric field (*i.e.* the negative gradient of the electrostatic potential) above the heterocycle. This latter parameter captures the electrostatic heterogeneity above the heterocycle and provides a more robust descriptor of the ability of a given heterocycle to electrostatically interact with the local dipole moment associated with the amide.[107] Given the large dipole associated with salt-bridges, one would expect similar molecular descriptors derived from electric fields and electrostatic potentials (ESPs) to be important in predicting the strength of salt-bridge stacking interactions.

Herein, we present accurate interaction energies of stacked dimers of an acetate-guanidinium salt-bridge (to model the Asp-Arg and Glu-Arg salt bridges prevalent in proteins) with a set of 63 pharmaceutically relevant heterocycles. We then use multivariate linear regression based on newly-developed heterocycle descriptors to develop simple predictive models for salt-bridge stacking energies based solely on the properties of the heterocycle that is effective across monocycles, bicycles, and tricycles. We also show that ESPs and electric fields can be used to rationalize observed trends in the interaction geometry, and present a set of guidelines for optimizing salt-bridge stacking interactions in ligand design.

Theoretical Methods

A set of 63 heterocycles (Schemes 3.1 and 3.2), including monocycles, bicycles and a limited number of tricycles, was selected to be representative of heterocyclic systems of interest in the design of biologically active ligands,[108-109] as well as their tautomers and regioisomers to provide a full picture of broad trends. Geometry

optimizations were done using ωB97X-D/def2-TZVP,[110-111] a range-separated and dispersion-corrected density functional method paired with a robust triple-ζ basis set. To ensure thorough sampling of local stacked energy minima, geometry optimizations commenced at 12 distinct starting points obtained by placing the heterocycle 3.5 Å above the plane of the salt-bridge either with the center of masses aligned or with a 1.5 Å shift. Six initial geometries were derived from each of these structures by rotating the heterocycle around the axis through the salt-bridge center of mass. Constraints were applied to ensure the heavy atoms of the salt-bridge and heterocycle remained in parallel planes. Following optimization, an RMSD cut-off of 0.4 Å was used to identify unique structures, resulting in a total of 255 local energy minima. For each heterocycle, there were between one and seven unique stacked geometries. Interaction energies, defined as the energy difference between the dimer and relaxed monomers, were computed at the DLPNO-CCSD(T)/cc-pVQZ[87-90, 112] level of theory for all 255 local minima. The agreement between these high-accuracy *ab initio* data and ωB97X-D/def2-TZVP predicted interactions energies (RMSE = 0.32 kcal mol^{-1} across 255 examples) supports the use of the latter method for geometry optimizations. Solvent corrections to these interaction energies were computed at the SMD-ωB97X-D/def2-TZVP level of theory in diethyl ether (ε = 4.24) as an approximate model of the dielectric environment of a protein.[113-115] DFT computations were performed using Gaussian 09,[94] while the DLPNO-CCSD(T) energies were computed using ORCA 4.0.1.[95] Zeroth-order symmetry adapted perturbation theory (SAPT0/jun-cc-pVDZ)[116-119] computations were done using Psi4.[120]



Scheme 3.1. Monocyclic heterocycles.

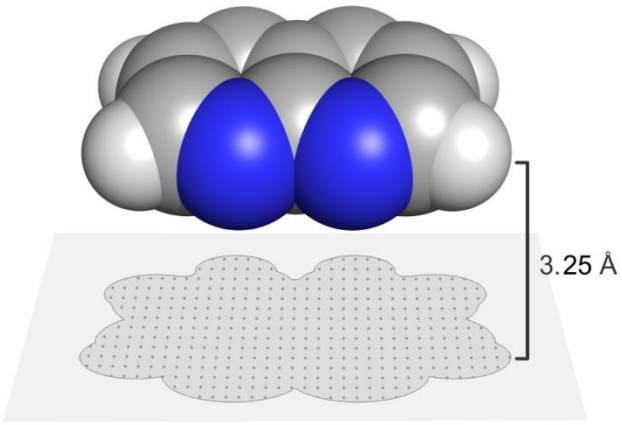
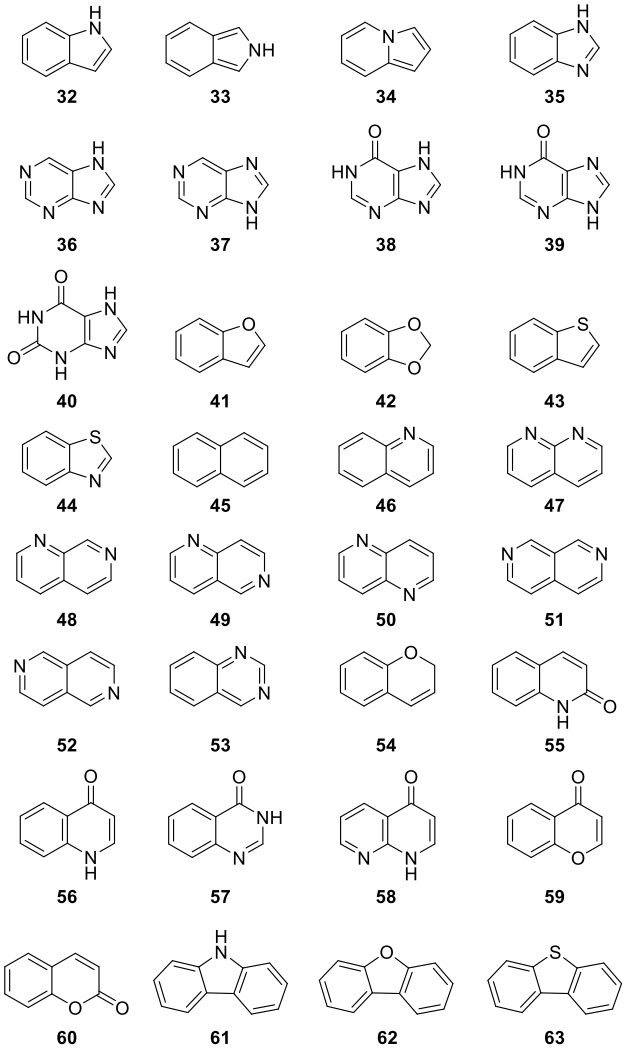


Figure 3.2. Electrostatic potential and electric field values were computed on a grid of points in a plane 3.25 Å from the plane of each heterocycle within the van der Waals projection of the heterocycle.



Scheme 3.2. Bicyclic and tricyclic heterocycles.

In order to devise robust predictive models of the strength of salt-bridge stacking interactions in both the gas phase and in solution, we developed eight molecular descriptors that capture features of the electrostatic environment in the plane above the heterocycles. Five of these descriptors are based on the electrostatic potential (ESP_{min}, ESP_{max}, ESP_{range}, ESP_{av}, and ESP_{σ}) and the remaining three are based on the electric field (F_{max}, F_{av}, and F_{σ}). Computed ESPs have been widely used to provide insight into a wide range of non-covalent interactions.[121-124] However, in most qualitative applications of ESPs the electrostatic potential is typically evaluated and plotted on an electron density

isosurface. In previous quantitative applications, the ESP was typically computed at one or a few specific points. Electric fields around heterocycles have also proved useful in providing a qualitative understanding of substituent effects in stacking interactions.[74, 106]

For these new descriptors, a uniform grid with spacing between points of 0.1 Å was defined as shown in Figure 3.2 and the ESP and the in-plane component of the electric field computed at each point. This region of space was chosen as the plane in which many stacking interactions occur, and so that the grid size is naturally appropriate for the size of the heterocycle.[125] These parameters were computed for all heterocycles in both the gas phase and in diethyl ether as a model of the dielectric environment of a protein binding site. Evaluating the ESP and electric field on this grid captures both local and global impacts of the heterocycle on the electrostatic environment via statistical measures of the ESP and field, giving the following descriptors:

- ESP_{min} – the minimum value of the ESP within the defined area
- ESP_{max} – The maximum value of the ESP within the defined area
- ESP_{range} – the difference between ESP_{max} and ESP_{min}
- ESP_{av} – The mean value of the ESP over all grid points
- ESP_{σ} – the standard deviation of the ESP over all grid points
- F_{max} – the maximum electric field magnitude within the defined area
- F_{av} – The mean of the electric field magnitude over all grid points
- F_{σ} – the standard deviation of the electric field magnitude over all grid points

These and 14 other descriptors were considered in the development of multivariate predictive models, which were determined through least squares regression

and cross-validated using L1O, 5-fold, and 10-fold analyses as implemented in scikit-learn version 17.1.[91]

Results and Discussion

A. Model Salt-Bridge Stacking Interactions

A preliminary model of salt-bridge stacking is shown in Figure 3.1A, in which one ring of a bicycle interacts with the anionic component of the salt-bridge via anion-π interactions while the other ring of the bicycle engages in a cation-π interaction with the cationic component. Conventional cation-π and anion-π interactions involving monoatomic cations and anions are dominated by electrostatic interactions,[71] with additional contributions from induction in the former case[126-128] and dispersion in the latter.[129-130] However, in the case of salt-bridge stacking interactions, the anion and cation are both delocalized over the respective π-systems; in such cases, one would expect considerable contributions from dispersion interactions, as found for stacking interactions involving two aromatic rings.[71, 76, 131]

Computed gas phase interaction energies for the global minimum energy stacked dimers of the model salt-bridge with the 63 heterocycles are listed in Table 3.1. Representative geometries are shown in Figure 3.3A. The average stacking distance is 3.1 Å across all of the global minimum dimers (see Appendix B, Figure B-1 for the distribution of inter-plane separations).

Table 1. Computed gas phase and in solution-phase interaction energies (E_{int}) as well as predicted interaction energies from equations (3.1) and (3.2), E_{int}^{pred}, for the global minimum energy stacked dimers of the model salt-bridge with the heterocycles in Schemes 3.1 and 3.2, in kcal mol^{-1}.

	Gas		Sol.			Gas		Sol.	
Het.	E_{int}	E_{int}^{pred}	E_{int}	E_{int}^{pred}	Het.	E_{int}	E_{int}^{pred}	E_{int}	E_{int}^{pred}
1	-8.8	-9.6	-2.8	-2.8	**33**	-13.4	-13.0	-5.4	-5.2
2	-9.5	-9.2	-3.4	-2.9	**34**	-11.9	-10.8	-5.1	-4.5
3	-11.3	-11.3	-3.4	-3.5	**35**	-13.9	-13.2	-5.1	-5.1
4	-10.1	-10.1	-2.4	-3.3	**36**	-15.7	-15.5	-5.6	-6.3
5	-13.8	-14.3	-4.2	-4.6	**37**	-12.3	-12.4	-4.0	-5.1
6	-12.3	-12.0	-5.4	-4.1	**38**	-13.5	-13.0	-4.5	-4.8
7	-5.8	-6.8	-2.3	-2.3	**39**	-15.7	-15.4	-5.6	-5.6
8	-9.0	-8.9	-4.4	-3.3	**40**	-13.2	-13.6	-4.7	-4.9
9	-13.6	-13.8	-5.5	-4.8	**41**	-10.2	-9.8	-4.5	-3.9
10	-5.9	-6.7	-1.9	-2.1	**42**	-10.5	-11.2	-4.8	-4.0
11	-8.7	-9.1	-3.5	-3.4	**43**	-10.6	-9.9	-4.6	-4.4
12	-6.8	-7.5	-2.6	-2.6	**44**	-10.4	-9.7	-5.0	-4.5
13	-9.5	-9.8	-4.2	-4.1	**45**	-10.5	-10.6	-4.3	-4.7
14	-6.7	-7.1	-2.6	-3.0	**46**	-11.7	-11.3	-5.2	-5.2
15	-8.3	-8.2	-3.3	-3.8	**47**	-14.1	-13.4	-6.3	-6.1
16	-7.6	-7.4	-3.1	-3.4	**48**	-12.3	-12.2	-6.1	-5.8
17	-10.0	-10.1	-4.8	-4.4	**49**	-11.8	-11.4	-5.5	-5.4
18	-10.9	-11.3	-2.8	-3.1	**50**	-10.5	-10.1	-5.0	-4.8
19	-6.8	-7.9	-2.5	-3.1	**51**	-10.0	-10.1	-4.9	-4.8
20	-8.4	-8.7	-3.2	-3.7	**52**	-9.9	-9.9	-4.8	-4.8
21	-11.0	-11.4	-5.0	-5.0	**53**	-11.0	-10.9	-4.7	-5.2
22	-8.5	-8.6	-2.9	-3.9	**54**	-12.5	-12.1	-5.6	-5.0
23	-6.5	-6.9	-3.4	-3.4	**55**	-11.1	-12.7	-4.5	-5.5
24	-6.9	-6.1	-2.4	-3.0	**56**	-16.1	-16.3	-6.9	-6.7
25	-10.4	-8.9	-4.7	-4.6	**57**	-12.2	-12.5	-4.9	-5.1
26	-12.6	-12.7	-6.2	-5.6	**58**	-13.2	-14.1	-5.2	-5.9
27	-12.3	-12.6	-6.2	-5.9	**59**	-12.7	-13.0	-5.7	-5.6
28	-10.6	-9.5	-5.4	-4.8	**60**	-12.5	-13.1	-5.6	-5.8
29	-8.6	-7.7	-5.0	-4.3	**61**	-14.0	-13.6	-6.0	-5.6
30	-12.2	-12.1	-4.6	-4.9	**62**	-10.9	-11.9	-4.9	-5.0
31	-15.0	-14.9	-5.4	-6.1	**63**	-12.0	-12.4	-5.7	-5.5
32	-13.4	-12.1	-5.4	-4.7					

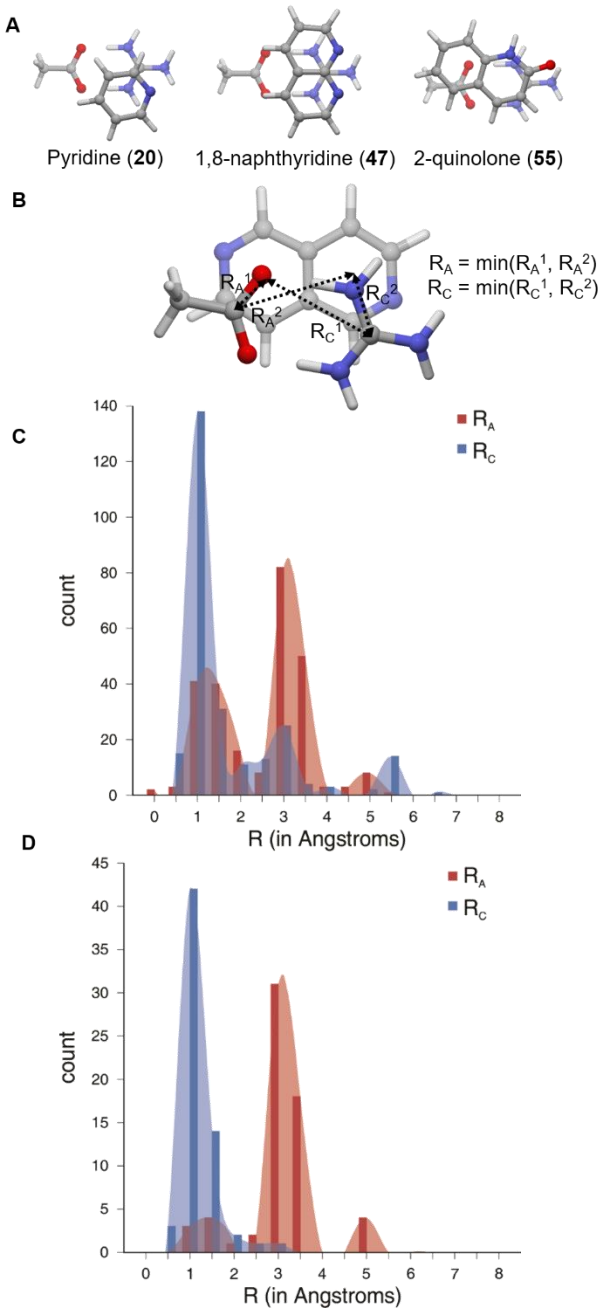


Figure 3.3. A) Representative global minimum energy stacked dimers of the model salt-bridge with selected heterocycles; B) definition of R_A and R_C, where R_A is the distance between the central carbon on the anionic fragment and the closest ring centroid projected into the heterocycle plane while R_C is defined similarly for the cationic fragment; C) distribution of R_A and R_C values for all local stacked energy minima; and D) distribution of R_A and R_C values for the global minimum energy stacked structures for each heterocycle.

The distribution of cation···ring centroid (R_C) and anion···ring centroid (R_A) distances (defined in Figure 3.3B) for all stacked local energy minima are plotted in Figure 3.3C. While the model depicted in Figure 3.1A would have both R_A and R_C close to zero, we see in Figure 3.3C, as well as in the representative examples in Figure 3.3A, that such a geometry is uncommon. Instead, the guanidinium distribution is clustered around 1.25 Å, which corresponds to geometries where the central carbon is located over the ring edge with one or two of the NH_2 groups positioned over the centroid of an aromatic ring. The carboxylate fragment shows a bimodal distribution, with one peak around 1.5 Å (corresponding to a similar configuration as the guanidinium ion, with the central carbon over the edge of the ring and a carboxylate oxygen over a ring centroid), while the second, larger peak is centered at 3.5 Å. This latter peak corresponds to dimers in which the carboxylate is positioned off to the side of the heterocycle. Interestingly, when we consider the R_C and R_A values for only the global minimum energy dimers (Figure 3.3D), the first peak in the carboxylate distribution is significantly diminished, indicating that the second configuration, where the heterocycle primarily interacts with the cationic fragment of the salt-bridge, tends to be more favorable. This is perhaps unsurprising, because cation-π interactions are much more likely to be energetically favorable than anion-π interactions.[71, 130] This is the case for both monocycles and bicycles, which adopt very similar configurations. For example, in the salt-bridge stacked complexes of pyridine (**20**) and 1,8-naphthyridine (**47**) the pyridine is in the same position as one of the two rings of naphthyridine, relative to the salt bridge (see Figure 3.3A). The majority of systems that deviate from this geometry are for heterocycles that contain an exocyclic carbonyl group. In these cases, specific interactions with the

carbonyl group override the more global effects, as seen for quinolone **55** (Figure 3.3A).[132]

B. Predictive Model of the Strength of Salt-Bridge Stacking Interactions

The interaction energies of the global minimum energy dimers span more than 10 kcal mol^{-1} across the 63 heterocycles studied (ranging from –16.1 to –5.8 kcal mol^{-1}); this range extends another 3 kcal mol^{-1} when all local minima are included. In order to understand what is driving this variation, contributions to the interaction energies were computed using SAPT0 (see Appendix B, Figure B-2).[133] The lack of correlation (R^2 = 0.01) between the total SAPT0 interaction energy (E_{SAPT0}) and the non-electrostatic components (*i.e.* $E_{non\text{-}elec} = E_{disp} + E_{ind} + E_{exch}$) indicates that the electrostatic component is essential to capture the trends in interaction energy. In other words, electrostatic interactions are the primary determinant of a strong salt-bridge stacking interaction. The relatively weak correlation (R^2 = 0.51) between E_{SAPT0} and $E_{non\text{-}disp}$ demonstrates that dispersion interactions are also important in capturing the variation in interaction energies.

In order to establish a simple predictive model of salt-bridge stacking interaction energies, we performed multi-parameter linear regressions based on 22 heterocycle descriptors. Given that SAPT0 analysis identified electrostatics as most important, followed by dispersion, we focused on descriptors capturing those features of the heterocycles. Among these 22 descriptors, there is no individual parameter that accurately predicts the interaction energy for the global minimum dimers. Indeed, there is a poor correlation between the heterocycle dipole moments and the total interaction energies (R^2 = 0.48), casting doubt on the viability of a simple dipole-based model of

these interactions. Considering only the electrostatic component of the interaction energies, ESP_{range} or ESP_{σ} provide the most robust predictions (R^2 = 0.90 and 0.85, respectively).[134] These correlations have a clear physical interpretation, because a strong positive ESP facilitates binding of the anionic component of the salt-bridge while a strong negative ESP is needed to bind the cation. Thus, a large variation in the ESP above the heterocycle should result in strong electrostatic stabilization of the salt-bridge. The dispersion component of the interaction energy can be described moderately well by the number of heavy atoms (N_{HA}),[135] volume, or polarizability (R^2 ranges from 0.72 to 0.81), which all reflect variations in the size of the heterocycle.

These descriptors can be combined using multivariate linear regression to give a linear fit that is able to predict the total interaction energies of the global minimum energy complexes. There are several combinations of electrostatic and dispersion descriptors that give rise to strong correlations (see Appendix B, Figure B-3 for additional fits considered). Although ESP_{range} provided the best fit for the electrostatic component of these interactions alone, the strongest correlation with the total interaction energy arises by combining ESP_{σ} with N_{HA},

$$E_{int}^{pred} = -1.53 ESP_{\sigma}^{gas} - 0.61 N_{HA} - 1.11 \qquad (3.1)$$

for which R^2 = 0.94, RMSE = 0.6 kcal mol^{-1}, and Q^2 = 0.93 (see Figure 3.4). This simple two-parameter model provides quantitative, robust predictions of salt-bridge stacking energies in the gas phase while maintaining a clear physical interpretation. Moreover, this fit applies equally well to monocyclic, bicyclic, and tricyclic heterocycles. Given the robustness of this fit combined with the simplicity of the underlying descriptors, equation

(3.1) should prove useful and reliable in identifying heterocycles that can potentially engage in strong stacking interactions with Asp/Glu-Arg salt-bridges.

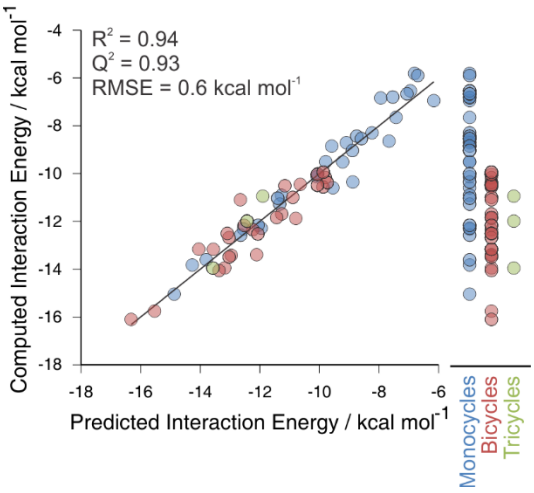


Figure 3.4. Correlation of computed gas phase interaction energies with predicted interaction energies from equation (1) for the global minimum energy salt-bridge stacked dimers. A strip plot of the computed interaction energies (in kcal mol^{-1}) grouped by heterocycle size is displayed to the right.

C. Specific Examples

Above, we showed that the strength of salt-bridge stacking interactions can be accurately predicted based on readily-computed heterocycle descriptors. However, in order to understand the preferred geometry of salt-bridge stacking, and to develop a set of guidelines for optimizing these interactions in the context of ligand design, it is helpful to examine series of related heterocycles. Although ESP_{σ} proved useful in predicting interaction energies, it is not well suited to describing geometries since it is a single statistical parameter characterizing the area above the heterocycle. To understand geometries, it is more useful to visualize the ESPs and electric fields in this region. Comparison of different diazanaphthalene regioisomers (Figure 3.5) allows us to look specifically at the way that the positions of heteroatoms in the heterocycle lead to

changes in the ESP and electric field, and how these changes impact the energy and geometry of the salt-bridge stacking interaction.

First, we consider 1,8-naphthyridine (**47**), above which there is a significant change in the ESP going from strongly positive above the CH side to strongly negative above the imino nitrogen side (see Figure 3.5A). This creates a strong electric field running perpendicular to the long axis of the heterocycle. In the global minimum energy stacked complex, the acetate is centered near the region of maximum positive ESP and the guanidinium is located in the region of maximum negative ESP. This aligns the dipole moment of the salt-bridge with the average electric field above the ring. The heteroatom arrangement in this naphthyridine isomer is ideal for simultaneously engaging with the salt-bridge via both dispersion and electrostatic interactions, and we see that **47** provides one of the strongest salt-bridge stacking interactions of -14.1 kcal mol^{-1}. In the case of 1,7-naphthyridine (**48**, Figure 3.5B), the direction of the electric field is rotated compared to 1,8-naphthyridine, and is no longer perpendicular to the long axis of the heterocycle. The salt-bridge rotates slightly in the corresponding dimer, with the net result of a weaker interaction energy of -12.3 kcal mol^{-1}. Continuing to 1,6-naphthyridine (**49**, Figure 3.5C), we can see further flattening of the ESP surface and weakening and rotation of the electric field relative to the long axis of the heterocycle. While the guanidinium ion in the corresponding dimer maintains two nitrogen atoms over the heterocyclic ring centroids, the relative position of the carboxylate has changed to align the salt-bridge dipole with the field. Finally, for 1,5-naphthyridine (**50**, Figure 3.5D) the ESP is fairly uniform, especially in the region directly over the heterocycle, with only a small anion stabilizing region located away from the heterocycle. The resulting small

electric field and reduced electrostatic stabilization leads to an interaction energy of only -10.5 kcal mol^{-1}. In these four dimers, the guanidinium is consistently positioned such that two nitrogen atoms are located above ring centroids and the salt-bridge dipole is aligned with the electric field above the ring. This tendency for the salt-bridge dipole moment to align with the average electric field is observed across all global minimum energy structures (see Appendix B, Figure B-4), particularly for heterocycles with strong electric fields.

The isoelectronic series of heterocycles in Figure 3.6 reveals similar trends for monocycles. As mentioned above, the salt-bridge stacked dimer with pyridine (**20**) is identical to one half of the 1,8-naphthyridine (**47**) dimer, with one of the hydrogen-bonding nitrogen atoms of the guanidinium positioned over the ring centroid and the salt-bridge dipole aligned with the electric field. The interaction energy for heterocycle **20** (-8.4 kcal mol^{-1}) is significantly smaller than that for **47** (-14.1 kcal mol^{-1}). This is expected, given both the decrease in heterocycle size and loss of a nitrogen going from **47** to **20**, which together result in reduced dispersion interactions and reduced variation in the ESP above the ring. For the diazines (**21** – **23**), we see that in all three cases the salt-bridge maintains the favored configuration with one nitrogen over the ring centroid and the anionic fragment oriented away from the two amino nitrogens of the diazine. For these three heterocycles, greater proximity of these two nitrogen atoms leads to stronger variations in the ESP above the ring. The result is that pyridazine (**21**) exhibits the strongest salt-bridge stacking (-11.0 kcal mol^{-1}), followed by pyrimidine (**22,** -8.5 kcal mol^{-1}), and finally pyrazine (**23**, -6.5 kcal mol^{-1}). Salt-bridge stacking interactions with the three triazine isomers (**24** – **26**) follow similar trends. The most favorable stacking is

exhibited by 1,2,3-triazine (**26**), in which the three adjacent nitrogen atoms enhance ESP_σ and result in a stronger stacking interaction compared to the other triazines in which the *N*-atoms are distributed more evenly around the ring.

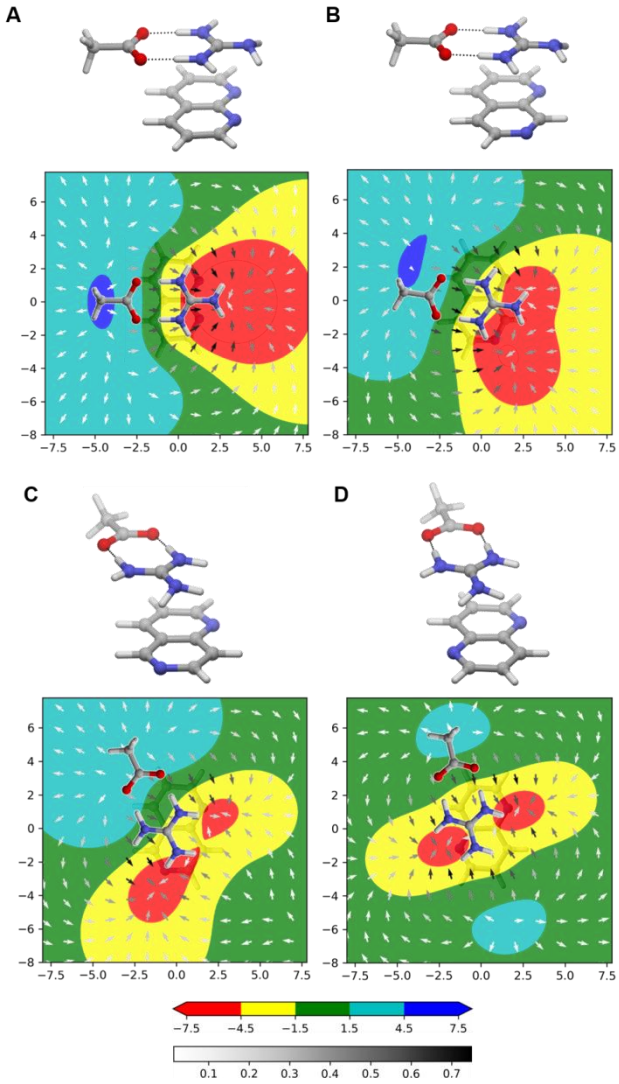


Figure 3.5. Global minimum energy stacked dimers and plots of the ESP and electric field arising from the heterocycle in the plane 3.25 Å above the heterocycle for A) 1,8-naphthyridine (**47**), B) 1,7-naphthyridine (**48**), C) 1,6-naphthyridine (**49**), and D) 1,5-naphthyridine (**50**). The colored contour plot shows the values of the electrostatic potential (in kcal mol^{-1}) while the arrows show the electric field direction and are shaded to show field strength (in kcal mol^{-1} D^{-1}).

Specific examples can also provide insight into the impact of increasing heterocycle size. For example, going from pyrrole (**1**) to indole (**32**) leads to a dramatic strengthening of the salt-bridge stacking interaction (E_{int} = -8.8 kcal mol^{-1} and -13.4 kcal mol^{-1} for **1** and **32**, respectively). Further expansion of the heterocycle to carbazole (**61**) leads to a more modest increase (E_{int} = -14.0 kcal mol^{-1}). For the related series of furan (**10**), benzofuran (**41**), and dibenzofuran (**62**), we see a similar trend with a large increase in interaction going from the monocycle to the bicycle followed by a more modest increase going from the bicycle to the tricycle. This is consistent with the idea that maximizing dispersion is an important feature, but is limited by the size of the salt-bridge itself. As the size of the heterocycle exceeds that of the salt-bridge, the impact of further appended rings is reduced.

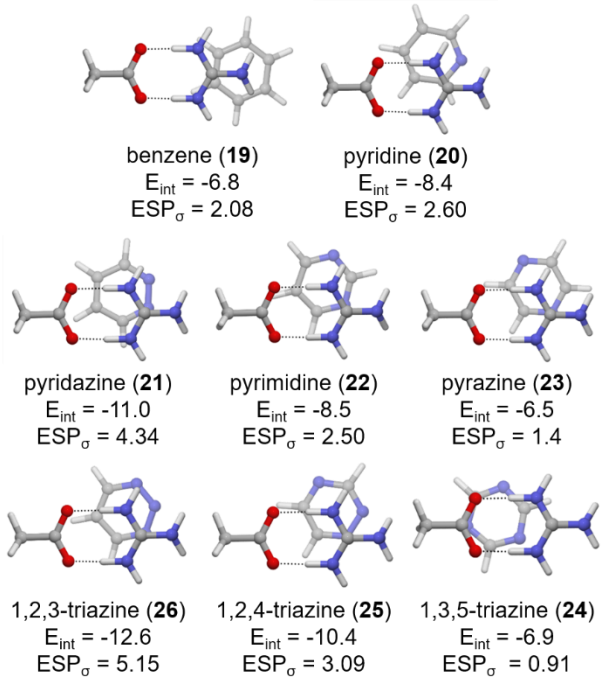


Figure 3.6. Global minimum energy dimers in the gas phase for selected heterocycles (E_{int} in kcal mol^{-1}, ESP_σ in kcal mol^{-1} e^{-1}).

It was noted above that the heterocycle dipole moment provides a relatively poor predictor of the strength of salt-bridge stacking interactions. There are several cases among the systems considered for which relative interaction energies are in conflict with expectations based solely on dipole moment. For example, naphthyridine **47** and quinolone **55** have nearly identical dipole moments, (4.26 and 4.18 D, respectively), but computed interaction energies differ by nearly 3 kcal mol^{-1} (-14.1 vs -11.1 kcal mol^{-1} for **47** and **55**, respectively). In this case, ESP_σ provides a sound predictor of interaction energies, with the heterocycle exhibiting a larger ESP_σ (**47**) resulting in stronger salt-bridge stacking interactions. These differences can be understood by looking at the ESPs and electric fields in the plane above the heterocycles (see Figure 3.7). For **55**, the strongest electric field is located between the N-H and C=O groups, which does not fall above the face of the heterocycle (Figure 3.7B). The ESP directly above the heterocycle is quite flat and the corresponding electric field small. This can be contrasted with naphthyridine **47,** where the electric field is maximized over the center of the heterocycle (Figure 3.7A). These subtle effects are reflected in the corresponding ESP_σ values. A second example that highlights the weakness of molecular dipole moments as a predictor of salt-bridge stacking is 2,6-naphthyridine (**52**) vs 2,7-naphthyridine (**51**). The small shift in heteroatom position causes the naphthyridine to go from having zero dipole moment for **52** to having a dipole moment of 1.88 D for **51**. However, there is no associated change in the stacking energies (E_{int} = -9.9 and -10.0 kcal mol^{-1}, respectively). The ESP_σ values for these two heterocycles are similar (1.75 vs 1.88 kcal mol^{-1} e^{-1}), in line with the small difference in interaction energies. In both cases (Figure 3.7C and D),

the large distance between the heteroatoms results in small variations in the ESP and weak electric fields.

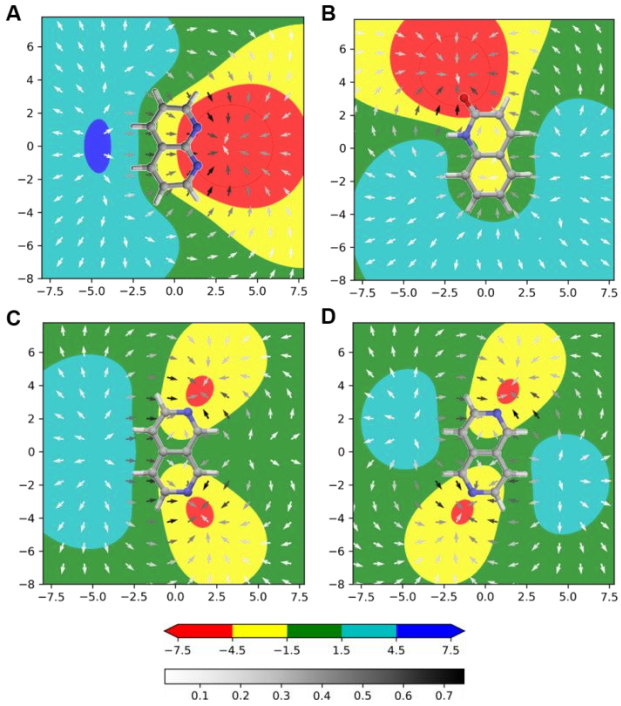


Figure 3.7. Electric Fields and ESPs in the plane 3.25Å above A) 1,8-naphthyridine (**47**), B) 2-quinolone (**55**), C) 2,-naphthyridine (**51**), and D) 2,6-naphthyridine (**52**). The colored contour plot shows the values of the electrostatic potential (in kcal mol^{-1}) while the arrows show the electric field direction and are shaded to show field strength (in kcal mol^{-1} D^{-1}).

D. Broad Trends and a Revised Qualitative Model

Based on these and other examples, combined with equation (3.1), we have developed the refined model of salt-bridge stacking interactions shown in Figure 3.1B. In this model, the cation sits directly over the bicycle with the anion off to the side and the salt-bridge dipole oriented perpendicular to the long axis of the bicycle. This new model provides a clear interpretation of the role of heteroatom and functional group placements in modulating the strength of salt-bridge stacking interactions. Given the tendency of the salt-bridge to orient orthogonal to the long axis of bicyclic heterocycles (see Appendix B,

Figure B-5), introduction of complementary substituents and heteroatoms along the long edges of bicyclic systems will have maximal impact on interaction energies as they lead to large variations in ESP over short length scales (leading to strong electric fields over the ring). This can be seen, for example, in the large change in interaction energy going from 1,8-naphthyridine (**47**) to 1,7-naphthyridine (**48**) compared to the smaller change going from 1,7-naphthyridine (**48**) to 1,6-naphthyridine (**49**). The largest ESP_σ values (and therefore salt-bridge stacking energies) arise when cation-binding groups (*e.g.* amino nitrogens) are clustered along one edge and anion-binding groups (imino nitrogens and exocyclic carbonyl groups) are clustered along the opposing edge. Having the cation and anion binding edges on the long edge of the heterocycle puts these two regions closer together, leading to a steeper change in the ESP and therefore stronger electric field. This also results in the cation and anion stabilizing regions of the ESP to be located at distances commensurate with the dimensions of the salt-bridge. Such groupings are seen in all of the observed strongest interacting dimers such as quinolone **56** (E_{int} = -16.1 kcal mol^{-1}) purine **36** (E_{int} = -15.7 kcal mol^{-1}), hypoxanthine **39** (E_{int} = -15.7 kcal mol^{-1}) and pyrimidin-2-one **31** (E_{int} = -15.0 kcal mol^{-1}).

E. Solvent Effects

The above analyses were based on gas phase interaction energies; however, we have also computed interaction energies for the same dimer geometries in diethyl ether (ε = 4.24) as a model of a typical protein environment.[113-115] While this does not account for the effect of heterocycle desolvation or any direct interactions with particular residues, these solution-phase interaction energies should provide a more realistic picture of salt-bridge stacking interactions in the dielectric environment of a protein. Solution phase

interactions energies for all 255 local energy minima are correlated with those in the gas phase (R^2 = 0.72, see Appendix B, Figure B-6), although the range of interaction energies (from +0.3 to -6.9 kcal mol^{-1}) is smaller. Notably, the geometric trends described above hold in solution, with the global minimum energy dimer for each heterocycle being the same in solution as in the gas phase for 84% of cases. Further, in the majority of cases where there is a discrepancy between the identity of the global minimum energy structure, the gas phase energy difference is less than 0.5 kcal mol^{-1}.

We developed a multivariate predictor of these solution phase interaction energies. While the two parameters used in equation 1 are not sufficient to accurately describe the solvent phase interaction energies (R^2 = 0.64, see Appendix B, Figure B-7), we were able to develop the following three-parameter predictive model of these data:

$$E_{int}^{pred} = -0.42ESP_{\sigma}^{Sol} + 0.03E_{LUMO}^{Sol} - 0.04\alpha_{zz} - 1.49 \qquad (3.2)$$

For this fit, R^2 = 0.83 and Q^2 = 0.82 (see Figure 3.8). While the correlation is not as strong as observed for the gas phase data, equation (3.2) is still capable of providing accurate predictions of salt-bridge stacking in a protein-like dielectric environment with a RMSE of only 0.5 kcal mol^{-1}. This predictor uses the same ESP-based descriptor (ESP_{σ}^{Sol}, computed in solution)[136] to account for the electrostatic effects as in the gas phase. However, these solution phase data required a more nuanced picture of dispersion interactions. In particular, rather than simply using the number of heavy atoms (N_{HA}), the solution-phase interaction energies requires the solution-phase LUMO energy (E_{LUMO}^{Sol}) and the polarizability perpendicular to the heterocycle plane (α_{zz}).

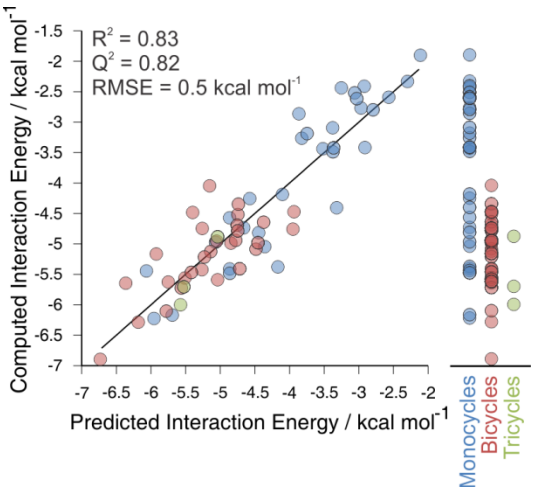


Figure 3.8. Correlation of computed interaction energies in solution with predicted interaction energies from equation (3.2) for the global minimum energy salt-bridge stacked dimers. A strip plot of the computed interaction energies (in kcal mol^{-1}) grouped by heterocycle size is displayed to the right.

Summary and Concluding Remarks

The many potential stacking partners in protein binding sites provide numerous opportunities to tune binding enthalpies. Stacking interactions involving Asp-Arg and Glu-Arg salt-bridges have been observed in of protein-ligand complexes, yet there has been relatively little effort to understand the means by which these interactions can be exploited in ligand design. We have presented accurate interaction energies for stacked dimers of a model salt-bridge with 63 heterocycles commonly found in pharmaceuticals in both the gas phase and in a polarizable continuum model of ether (mimicking a protein-like dielectric environment). The strength of these interactions can be tuned over a 10 kcal mol^{-1} range in the gas phase and 5 kcal mol^{-1} in ether, highlighting an underappreciated means of tailoring protein-ligand binding. While the range of interaction energies is reduced in solvent, the overall trends observed in the gas phase data hold. Further, even though this simple solvent model does not account for all the factors that influence binding in real systems, harnessing salt-bridge stacking interactions

can still lead to significantly improved ligand binding. The gas phase salt-bridge stacking interactions can be accurately predicted by a simple two-parameter linear model based on a new electrostatic potential derived heterocycle descriptor (ESP_σ) combined with the number of heterocycle heavy atoms (equation 3.1), whereas the solution-phase data requires a slightly more complex, three-parameter model (equation 3.2). These simple models provide physical insight into the strength of these interactions while examination of the ESP and electric field above the heterocycles can explain the preferred orientation. Moreover, these new ESP and electric-field based heterocycle descriptors should prove useful in the development of predictive models of other stacking interactions involving heterocycles.

Heterocycles that interact most strongly with salt-bridges have a number of common features (see Figure 3.1B). First, they are generally substituted bicycles with cation stabilizing groups (exocyclic carbonyls and imino nitrogens) clustered along one long edge and anion stabilizing groups (*e.g.* amino nitrogens) clustered along the other. This leads to large variation in the electrostatic potential and a large electric field across the short axis of the heterocycle. This, in turn, provides optimal electrostatic stabilization of the salt-bridge in a geometry that is compatible with strong dispersion interactions. In highly favorable salt-bridge stacking interactions, the cation sits over the arene (with nitrogen atoms over the ring centers), the anion sits off to the side, and the salt-bridge dipole is aligned with the electric field over the heterocycle (see Figure 2.1B). This refined view of salt-bridge stacking interactions, combined with the predictive models in equations (3.1) and (3.2), should prove useful in choosing heterocycle fragments that maximize these interactions in drug binding sites.

CHAPTER 4

PREDICTING THE STRENGTH OF STACKING INTERACTIONS BETWEEN HETEROCYCLES AND AROMATIC AMINO ACID SIDE CHAINS [c]

[c] Bootsma, A. N, Doney, A. C. and S. E. Wheeler. Submitted to J. Am. Chem. Soc.

Abstract

Despite the ubiquity of stacking interactions between heterocycles and aromatic amino acids in biological systems, our ability to predict their strength, even qualitatively, is limited. Based on rigorous *ab initio* data, we have devised a simple predictive model of the strength of stacking interactions between heterocycles commonly found in biologically active molecules and the amino acid side chains Phe, Tyr, and Trp. This model provides rapid predictions of the stacking ability of a given heterocycle based on readily-computed heterocycle descriptors. We show that the values of these descriptors, and therefore the strength of stacking interactions with aromatic amino acid side chains, follow simple predictable trends and can be modulated by changing the number and distribution of heteroatoms within the heterocycle. This provides a simple conceptual model for understanding stacking interactions in protein binding sites and optimizing inhibitor binding in drug design.

Introduction

Arene stacking interactions, broadly defined as approximately parallel face-to-face interections between aromatic rings,[70, 137-138] impact everything from the binding of ligands by proteins to the solid-state packing of organoelectronic materials and the performance of asymmetric catalysts.[59-60, 96-97, 139-141] By understanding the factors that impact the strength of these interactions, we can explain their role in chemical and biochemical systems and exploit their power in the context of design. Stacking interactions between heterocycles and the aromatic amino acid side chains Phe, Tyr, and Trp are particularly important in biological systems,[142-144] contributing to the binding of artificial inhibitors, natural substrates, co-factors, and nucleic acids by proteins (see Fig

4.1).[145-150] Changes in the number and distribution of heteroatoms within such heterocyclic systems can modulate their ability to stack with binding site aromatic residues, tuning their overall binding over a broad range.[27]

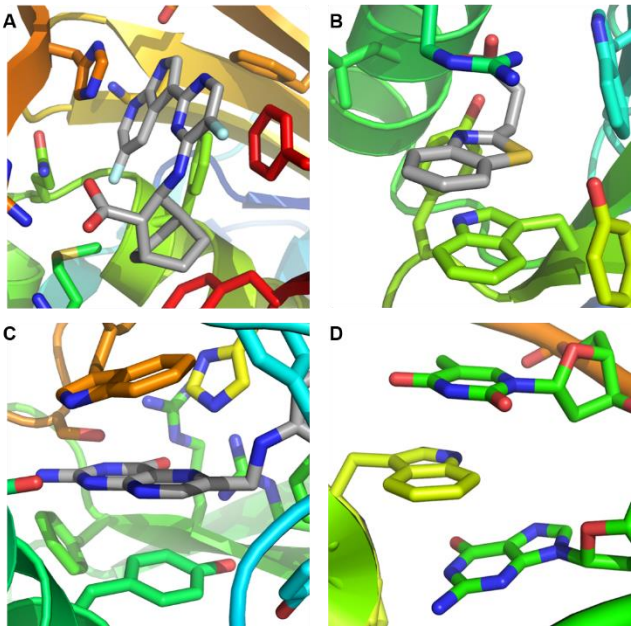


Figure 4.1. Examples of stacking interactions between heterocycles and aromatic amino acid side chains in biological systems. A. Inhibitor VX-787 (grey) bound to the influenza A viral polymerase (PDB: 4P1U); B. A quinazolinone-based inhibitor (grey) bound to a zinc-finger ubiquitin binding domain (PDB: 6CEF); C. Folate (grey) bound to the human folate receptor (PDB: 4KMZ). D. Single stranded DNA (green) bound to human TDP-41 (PDB:4IUF).

The favorable stacking between aromatic rings has long been recognized, and Hunter and Sanders[151] provided a foundational conceptual model of the factors that impact the strength and geometry of these interactions. Over the last 15 years, high-accuracy quantum chemical studies and new experiments have provided a more nuanced view of these interactions, particularly with regard to the impact of substituents.[69, 72, 76, 152-159] However, despite progress in the area of heterocycle stacking,[160-169] the factors that impact how strongly a given heterocycle will stack are still not completely understood. For instance, we currently have no means, short of expensive quantum chemical

computations, of ranking a set of heterocycles with regard to their ability to stack with aromatic amino acid side chains, even qualitatively. This limits our ability to understand stacking interactions in many biological contexts and hamstrings efforts at structure based drug design.

In 2009, Hohenstein and Sherrill[164] published high-accuracy gas-phase computations of the pyridine-benzene dimer, showing that electrostatic interactions result in a 0.5 kcal mol^{-1} enhancement in the stacking interaction compared to the benzene dimer. Subsequently, Stahl and co-workers[170] published a medicinal chemist's guide to molecular interactions based on analyses of crystal structure data and representative drug binding sites. This included the general advice that stacking interactions of electron-deficient rings are generally preferred over stacking of electron-rich rings and noted that the preferred orientation of stacked arenes can often be rationalized by the alignment of molecular dipoles or partial atomic charges.[96] In a different context, Corminboeuf and co-workers[171] showed that π-electron depletion correlates with the strength of π-stacking interactions, lending further support to the notion that electron-deficient arenes stack more effectively.

There have been several efforts to distill our understanding of heterocycle stacking into predictive models based on simple molecular descriptors. For instance, Huber and co-workers[30] demonstrated that molecular dipole moments are correlated with the strength of stacking interactions of monocyclic heterocycles with benzene, suggesting that the strength of stacking interactions can be predicted without resorting to expensive quantum chemistry computations. However, subsequent work by An, *et al.*[31] on stacking interactions between a broader set of biologically relevant heterocycles and 9-

methyladenine showed that dipole moments alone do not capture trends in stacking interactions for bicyclic and larger heterocycles. Thus, it appears that the development of a predictive model of the strength of stacking interactions applicable to the diverse heterocycles appearing in biologically active molecules will require more advanced molecular descriptors.

We have demonstrated the utility of newly developed heterocycle descriptors[172-173] in predictive models of non-covalent interactions relevant to biological systems. For instance, we were able to obtain robust predictions of stacking interactions between Asp-Arg salt-bridges and heterocycles commonly found in pharmaceuticals[173] using simple combinations of the our recently introduced heterocycle descriptors derived from the electrostatic potential (ESP) and electric field near the heterocycle. Here, we use these descriptors to devise a predictive model of the strength of stacking interactions between heterocycles found in biologically active molecules and Phe, Tyr, and Trp side chains, which informs the development of a simple conceptual framework for understanding these interactions.

Results and Discussion

We systematically searched for low-lying stacked dimers of toluene, *p*-methylphenol, and 3-methylindole (as models of Phe, Tyr, and Trp side chains, respectively; see Chart 4.1) with 47 heterocycles that are representative of those in biologically active molecules (see Charts 4.1 and 4.2).[77, 79] The gas-phase interaction energies for all unique stacked dimers were then computed using high-accuracy *ab initio* methods. Interaction energies for the global minimum energy stacked dimers of each heterocycle with Phe, Tyr, and Trp are listed in Table 4.1. Figure 4.2 shows geometries of

selected dimers. The computed interaction energies span more than 10 kcal mol^{-1} across all heterocycles considered. This includes a range of 8.9 kcal mol^{-1} across the monocycles, from –3.6 kcal mol^{-1} for pyrazole (**8**) stacked with Phe to–12.5 kcal mol^{-1} for pyrimidin-2-one (**27**) stacked with Trp. As for the bicyclic systems, the range is 8.7 kcal mol^{-1}, from–7.1 kcal mol^{-1} for isindole (**29**) stacked with Phe to –15.8 kcal mol^{-1} for 3,9-dihydro-purine-2,6-dione (**38**) stacked with Trp. Stacking interactions of heterocycles with Tyr are, on average, ~10% stronger than those with Phe, while stacking interactions with Trp are ~40% stronger. Apart from this broad trend, there is a considerable spread in predicted interaction energies for each amino acid side chain. This highlights the wide range over which heterocycle stacking interactions with a given aromatic amino acid can vary while also revealing the key means of tuning stacking interactions in the context of drug design.[27]

In order to understand the trends in stacking interaction energies across this set of heterocycles, these dimers were first analyzed using symmetry-adapted perturbation theory (SAPT).[116-119] SAPT enables the decomposition of interaction energies into electrostatic, dispersion, exchange-repulsion, and induction effects, and can be used to identify the effects that drive trends in interaction energies.[174] Overall, these data show that the stacking abilities of the heterocycles in Charts 4.1 and 4.2 are driven by dispersion and electrostatic effects (see Appendix C, Figure C-1). This is in accord with the previous results of Hohenstein and Sherrill[164] for the stacked benzene-pyridine dimer.

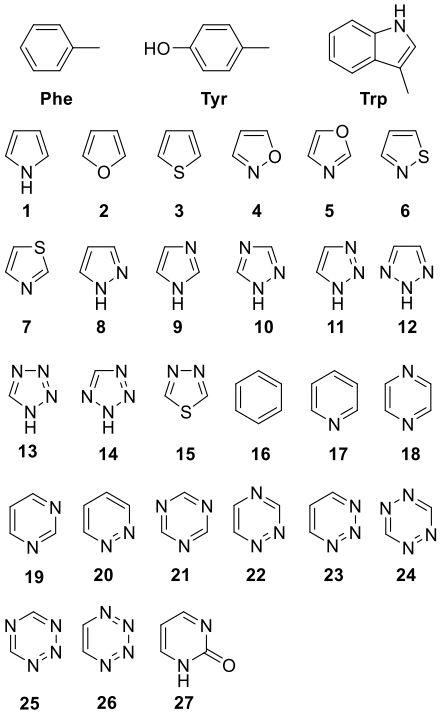


Chart 4.1. Set of monocyclic heterocycles considered along with the model amino acid side chains

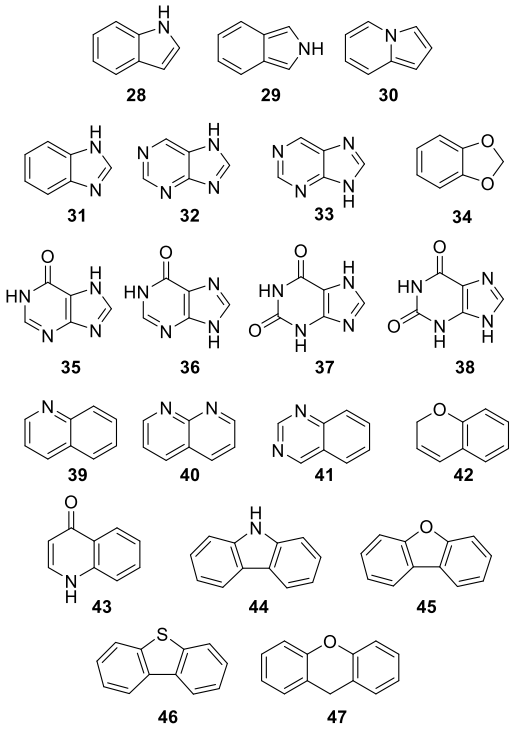


Chart 4.2. Set of bicyclic and tricyclic heterocycles considered

Based on this insight, we sought to develop a predictive model of the stacking energy between each of these heterocycles and the aromatic amino acids by applying multivariate linear regression (MLR) methods. This fit used simple descriptors of the heterocycles and amino acid side chains that capture their ability to engage in dispersion and electrostatic interactions. Initial fitting of the data for Phe, Tyr, and Trp separately revealed that the coefficients in each fit were proportional to the size of the amino acid side chain. This motivated the form of equation 4.1, which depends on only three heterocycle descriptors and one descriptor of the amino acid and was fit to the global minimum energy dimers of heterocycles **1**-**47** with Phe, Tyr, and Trp:

$$\Delta E_{pred} = N_{HA}^{AA}(-0.01ESP_{range}^{Het} - 0.04ESP_{mean}^{Het} - 0.10N_{HA}^{Het}) - 1.25 \quad (4.1)$$

In this fit, N_{HA} refers to the heavy atom count of either the amino acid (N_{HA}^{AA}) or heterocycle (N_{HA}^{Het}) while ESP_{mean}^{Het} and ESP_{range}^{Het} are descriptors[173] based on the mean value and range of the ESP within the van der Waals projection of the heterocycle in a plane 3.25 Å from the heterocycle. The predicted interaction energies from equation 4.1 are plotted against the DLPNO-CCSD(T) interaction energies in Figure 4.3. This fit is robust, indicating that the maximum potential stacking interaction of a given heterocycle with Phe, Tyr, or Trp can be reliably predicted without resorting to expensive QM computations of stacked dimers. Instead, one needs to simply evaluate ESP_{mean}^{Het} and ESP_{range}^{Het}, which only require simple computations on the isolated heterocycles.

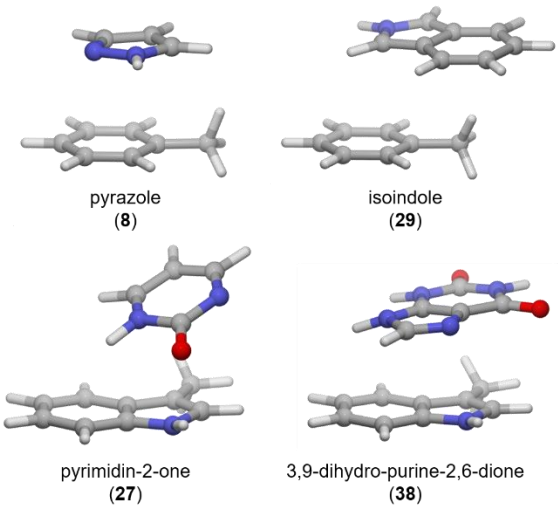


Figure 4.2. Stacked dimers of selected heterocycles with Phe and Trp.

The only descriptor of the amino acid side chain in equation 1 is the heavy atom count,[135] N_{HA}^{AA}, which typically correlates with the ability of a given system to engage in dispersion-driven interactions. However, since there are only three amino acids, this parameter should not be over-interpreted. It may be that the ability of these amino acid side chains to engage in dispersion interactions is the only important factor; however, it could be that N_{HA}^{AA} (7 for Phe, 8 for Tyr, and 10 For Trp) simply encodes the general ability of these three arenes to engage in stacking interactions due to other factors. Essentially, N_{HA}^{AA} provides a scaling factor of the strength of stacking interactions that reflects the general observation that stacking interactions with Tyr and Trp are, on average, ~10% and ~40% stronger than with Phe, respectively.

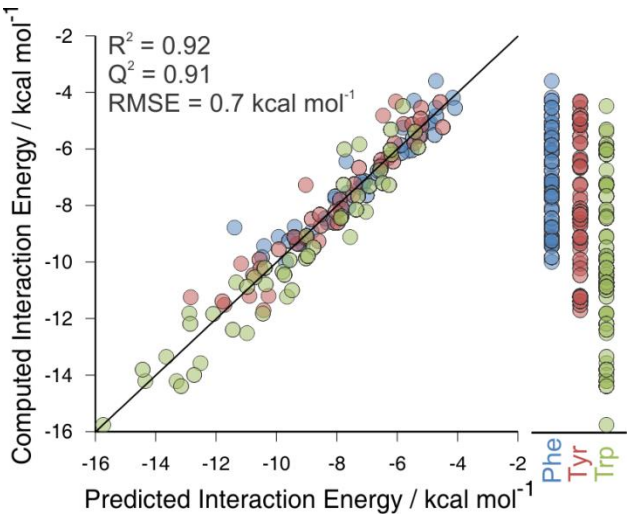


Figure 4.3. Scatter plot of computed DLPNO-CCSD(T) interaction energies vs predicted interaction energies from equation 4.1 for global minimum energy stacked dimers of heterocycles **1**-**47** with Phe, Tyr, and Trp side chains. The strip plot on the right shows the same computed interaction energies grouped by amino acid.

The heterocycle descriptors appearing in equation 4.1 (ESP_{range}^{Het}, ESP_{mean}^{Het}, and N_{HA}^{Het}), on the other hand, confirm a prominent role of electrostatic effects and dispersion interactions, in accord with the SAPT data. Overall, each additional heavy atom in the heterocycle leads to a 0.7, 0.8 and 1.1 kcal mol^{-1} increases in the stacking interaction with Phe, Tyr, and Trp, respectively. With regard to electrostatics, stacking interactions are maximized in heterocycles with a large positive mean ESP value and a large overall range in ESP values in the plane 3.25 Å from the heterocycle. More precisely, equation 1 indicates that a 1 kcal mol^{-1} increase in ESP_{range}^{Het} corresponds to a 0.3-0.4 kcal mol^{-1} increase in stacking energies, while a 1 kcal mol^{-1} increase in ESP_{mean}^{Het} translates into ~1 kcal mol^{-1} increased stacking energy. It should be noted that ESP_{range}^{Het} is more sensitive to the introduction of heteroatoms and other functionality than is ESP_{mean}^{Het} (*vide infra*), and modulating the value of both of these descriptors provide a means of widely tuning the strength of stacking interactions.

We can consider specific examples of heterocycles in order to gain a more intuitive understanding of the heterocycle descriptors appearing in equation 1 and the impact of changes in the number and distribution of heteroatoms on stacking interactions (see Table 4.1).

Table 4.1. Interaction energies for global minimum energy stacked dimers of heterocycles 1-47 with Phe, Tyr, and Trp side chains along with heterocycle descriptors, all in kcal mol^{-1}.[a]

Het	ESP^{Het}_{mean}	ESP^{Het}_{range}	E^{Phe}_{int}	E^{Tyr}_{int}	E^{Trp}_{int}	Het	ESP^{Het}_{mean}	ESP^{Het}_{range}	E^{Phe}_{int}	E^{Tyr}_{int}	E^{Trp}_{int}
1	-5.9	14.7	–	–	-4.5	**25**	3.0	11.2	-7.1	-7.8	-9.9
2	-3.9	6.3	-4.2	-4.3	-5.4	**26**	3.1	16.8	-8.0	-8.7	-10.8
3	-4.5	7.0	-4.5	-5.2	-5.9	**27**	0.1	22.4	-7.8	-7.3	-12.5
4	-1.8	11.5	-5.5	-5.1	-7.3	**28**	-6.2	15.5	-7.3	-8.6	-11.2
5	-1.8	9.0	-5.1	-5.1	-6.7	**29**	-6.7	17.7	-7.1	-8.2	-10.4
6	-2.4	10.1	-5.1	-5.8	-7.2	**30**	-5.1	11.4	-7.6	-8.1	-11.0
7	-2.5	8.8	-5.5	-5.5	-7.3	**31**	-4.5	17.0	-6.4	-9.3	-11.8
8	-4.0	12.2	-3.6	-4.5	-5.3	**32**	-0.7	21.8	-9.4	-11.2	-13.6
9	-3.7	17.7	–	–	-8.2	**33**	-1.1	14.9	-8.4	-9.1	-12.4
10	-1.7	13.8	-4.3	-4.3	-5.8	**34**	-3.0	12.4	-7.6	-8.3	-10.2
11	-1.8	17.8	–	–	-6.0	**35**	-0.3	15.3	-9.4	-11.7	-14.0
12	-2.1	7.0	-4.8	-5.1	-6.2	**36**	-1.0	20.4	-9.2	-10.5	-14.4
13	0.5	20.9	–	–	-9.1	**37**	1.1	16.4	-9.4	-11.5	-14.2
14	0.0	13.0	-5.2	-4.8	-7.3	**38**	0.9	27.1	-8.8	-11.2	-15.8
15	-0.5	12.7	-6.1	-6.3	-9.1	**39**	-3.9	10.2	-8.1	-8.5	-10.5
16	-4.7	7.4	-4.5	-4.9	-6.0	**40**	-2.3	16.1	-8.9	-9.5	-11.8
17	-2.8	10.0	-5.6	-5.8	-7.7	**41**	-1.8	9.1	-8.6	-9.3	-10.7
18	-0.9	5.4	-6.0	-6.4	-8.1	**42**	-4.2	12.8	-7.7	-9.1	-10.9
19	-0.8	8.8	-6.1	-6.4	-8.4	**43**	-3.2	22.8	-9.1	-10.1	-13.3
20	-0.7	15.4	-6.5	-6.7	-9.5	**44**	-5.8	13.2	-10.0	-11.2	-14.2

21	1.1	3.8	-5.9	-6.6	-8.4	**45**	-4.6	6.9	-8.8	-9.9	-11.8
22	1.1	12.3	-7.0	-7.6	-9.9	**46**	-5.0	7.6	-9.2	-10.2	-12.2
23	1.2	17.6	-7.3	-8.6	-10.3	**47**	-4.4	10.6	-9.8	-11.4	-13.8
24	2.9	7.0	-6.8	-7.2	-9.8			**Mean**	**-7.1**	**-7.8**	**-9.8**

[a] Dashes indicate that no stacked local energy minima were located.

For instance, the role of heterocycle size (as captured by N_{HA}^{Het}) is demonstrated by stacking interactions of pyrrole (**1**), indole (**28**), and carbazole (**44**). These three heterocycles have similar values for ESP_{range}^{Het} and ESP_{mean}^{Het} but vary in size. Computed interaction energies for these three heterocycles stacked with Trp, for example, follow the trend in size, going from –4.5 kcal mol^{-1} for **1**, to –11.2 kcal mol^{-1} for **28**, finally to –14.2 kcal mol^{-1} for **44**.

Examination of the values for the electrostatic descriptors ESP_{mean}^{Het} and ESP_{range}^{Het} across the heterocycles reveals clear trends that can be distilled into simple rules for their estimation, and thus the strength of stacking interactions. For example, ESP_{mean}^{Het} depends primarily on the number of S, O, imino N (-N=), amino N (-NH-), and carbonyl groups present in the heterocycle. Overall, each S, O, imino N, and carbonyl increases ESP_{mean}^{Het} by 0.5, 1.0, 2.0, and 3.5 kcal mol^{-1}, respectively. In contrast, each amino nitrogen decreases ESP_{mean}^{Het} by 1 kcal mol^{-1}. These effects are additive, as seen through the consecutive addition of imino nitrogens to benzene (**16**) to give pyridine (**17**), pyrimidine (**19**), and 1,2,4-triazine (**22**). The ESP_{mean}^{Het} values for these rings systematically increases from -4.7 to 1.1 kcal mol^{-1}, accompanied by an improvement in interaction energy with Trp from -6.0 kcal mol^{-1} for benzene to -9.9 kcal mol^{-1} for 1,2,4-triazine. This dependence of ESP_{mean}^{Het} on the number of these different heteroatoms is the origin of the general observation that electron deficient heterocycles (*e.g.* those containing imino

nitrogens) tend to engage in stronger stacking interactions,[96, 170-171] and ESP_{mean}^{Het} provides a readily-computed means of quantifying the degree of electron deficient character for a given heterocycle.

However, these changes in ESP_{mean}^{Het} do not fully capture variations in the strength of stacking interactions with heterocycles. For example, despite the same number of heteroatoms in pyrazine (**18**), pyrimidine (**19**) and pyridazine (**20**), and their correspondingly similar ESP_{mean}^{Het} values (–0.8 ± 0.1 kcal mol^{-1}), their stacking interactions with Trp vary from –8.1 to -9.5 kcal mol^{-1}. This can be attributed to the variation in ESP_{range}^{Het} across this series.

Unlike ESP_{mean}^{Het} which depends only on the number of heteroatoms, the value of ESP_{range}^{Het} depends on both the number of each of these heteroatoms and functional groups as well as their relative position within the heterocycle. Qualitatively, when a pair of 'like' heteroatoms (S, O, imino N, and C=O vs amino N) are on nearby positions, there is a cooperative impact on ESP_{range}^{Het}. For instance, whereas the addition of a single imino nitrogen to form pyridine (**17**) from benzene (**16**) shifts ESP_{range}^{Het} by 2.7 kcal mol^{-1}, the introduction of a second *N* to form pyrimidine (**20**) has nearly twice the effect. On the other hand, when like heteroatoms are on opposite sides of a ring, their impact on ESP_{range}^{Het} cancels to a large degree. This can be seen by comparing pyrimidine (**19**) and pyrazine (**18**), which both have smaller values of ESP_{range}^{Het} than pyridine. Indeed, ESP_{range}^{Het} for pyrazine is smaller than that for benzene. Similar trends hold for heterocycles containing any combination of S, O, amino N, and C=O groups. For 'unlike' heteroatoms, the opposite trend occurs, with the impact of adjacent heteroatoms canceling and opposing heteroatoms exhibit positive cooperativity. This can be

understood by recalling that amino and imino nitrogens have opposite impacts on the ESP, as demonstrated by the opposite sign of their impact on ESP_{mean}^{Het}. For example, for pyrazole (**8**), which has amino and imino nitrogens at adjacent positions, the value of ESP_{range}^{Het} of 12.2 kcal mol^{-1} is lower than that of pyrrole (**1**, ESP_{range}^{Het} = 14.7 kcal mol^{-1}). At the same time, the 1,3 amino-imino nitrogen pair in imidazole (**9**) leads to an ESP_{range}^{Het} value of 17.7 kcal mol^{-1}. Thus, ESP_{range}^{Het} is maximized when like heteroatoms and functional groups are grouped together one side of the ring and the 'unlike' types on the opposite side of the ring. For example, the strongest stacking interactions is exhibited by 3,9-dihydro-purine-2,6-dione (**38**) stacked with Trp (see Figure 4.2). In this case, the presence of two C=O groups including one adjacent to an imino nitrogen overwhelms the modest effects of the three amino nitrogens, leading to a substantial value of ESP_{range}^{Het} , a high ESP_{mean}^{Het}, and very strong stacking interaction. Other examples of systems that stack strongly have similar distributions of heteroatoms and carbonyl groups.

While we have focused on the importance of ESP_{range}^{Het} in determining stacking interactions, it should be noted that molecular dipole moments (μ) provide similar information as ESP_{range}^{Het}. Indeed, a functional form analogous to equation 1 featuring μ instead of ESP_{range}^{Het} provides a reasonable predictor of these stacking interactions, but with a reduced R^2 value (see Appendix C, Figure C-3). However, ESP_{range}^{Het} is arguably a more versatile heterocycle descriptor since it can capture differences in electrostatic character among rings with zero net dipole (*e.g.* benzene, pyrazine, and 1,3,5-triazine).[172]

Conclusions

We have developed a simple predictive model (equation 4.1) of the maximum possible stacking interaction between heterocycles found in biologically active molecules

and the amino acid side chains Phe, Tyr, and Trp based on readily computed heterocycle descriptors. This model depends on the size of the heterocycle (as captured by the heavy-atom count),[135] as well as the electrostatic descriptors ESP_{range}^{Het} and ESP_{max}^{Het}.[173] These latter two descriptors follow simple, predictable trends and can be modulated by varying the number and distribution of heteroatoms and other functionality within the heterocycle. This provides a clear conceptual framework for understanding trends in stacking interactions between heterocycles and aromatic amino acid side chains and designing heterocycles that maximize such interactions. Overall, stacking interactions with Phe, Tyr, and Trp are enhanced by:[175]

1. Increasing the numbers of S, O, imino N, and carbonyl groups, with the size of this effect increasing across this series;
2. Grouping 'like' heteroatoms (S, O, imino N, carbonyl vs. amino N) on opposing sides of the heterocycle.

Using these rules, heterocycles can be qualitatively ranked with regard to their ability to stack with Phe, Tyr, and Trp. For more quantitative predictions, the descriptors appearing in equation 4.1 can be rapidly computed using widely available computational tools. This new view of stacking interactions between heterocycles and aromatic amino acid side chains should facilitate both a greater understanding of stacking in biological systems and provide updated principles for maximizing stacking interactions in protein binding sites.

Computational Section

Local energy minima were identified by systematically sampling six (or 12, for non-symmetric heterocycles) orientations of the heterocycle at each of nine initial points

located in a 3 x 3 grid 3.6 Å above the heterocycle (fewer for Phe due to symmetry). Each of these starting configurations was optimized in the gas phase at the B97D/def2-TZVPP[111, 176-177] level of theory. This dispersion-corrected DFT functional paired with a triple-ζ basis set should provide reliable geometries for these stacked dimers.[176] The resulting energy minima were then classified as either stacked or edge-to-face (T-shaped) geometries using the criteria from Ref 31. For each heterocycle/amino acid combination, there were up to 24 unique stacked energy minima (based on an RMSD cutoff of 0.4 Å). For eight heterocycle/amino acid combinations,[178] no stacked energy minima were located. DLPNO-CCSD(T)/cc-pVQZ single point energies[87-90, 112] were computed for each unique stacked energy minimum to obtain highly-accurate interaction energies (*i.e.* the energy difference between the optimized dimer and corresponding optimized monomers). These *ab initio* interaction energies are converged to within 0.1 kcal mol^{-1} of the complete basis set (CBS) limit (see Appendix C, Table C-1). Solution-phase interaction energies (in diethyl ether) were computed by combining the gas-phase DLPNO-CCSD(T) energies with solvation corrections computed at the ωB97X-D/def2-TZVPP level of theory. While the above focus is on the global minimum energy stacked dimer for each heterocycle with the three amino acid side chain models, data for all local stacked energy minima are available in the SI. All DFT optimizations were performed using Gaussian09,[94] while Orca 4.0[95] was used for the DLPNO-CCSD(T) single point energies. SAPT computations were performed at the SAPT0/jun-cc-pVDZ level of theory[116-119, 179] using Psi4.[120]

CHAPTER 5

RAPID EVALUATION OF ELECTROSTATIC POTENTIAL BASED HETEROCYCLE DESCRIPTORS: CONVERTING SMILES TO STACKING INTERACTION ENERGIES [d]

[d] Bootsma, A. N., and S. E. Wheeler. In preparation for submission to J. Chem. Inf. Model.

Abstract

Predicting the strength of stacking interactions involving heterocycles is vital for a number of fields, including drug design. While quantum chemical computations can provide accurate stacking interaction energies, these come at a steep computational cost. Previously, we developed a quantitative predictive model of stacking interactions between drug-like heterocycles and the aromatic amino acids Phe, Tyr, and Trp. This model depends on heterocycle descriptors derived from the electrostatic potential computed in a plane near the heterocycle. Herein, we show that these descriptors can be evaluated directly from the atom connectivity of the heterocycle, providing a means of rapidly evaluating both the descriptors and potential for a given heterocycle to engage in stacking interactions without resorting to any quantum chemical computations. This enables the conversion of simple molecular representations (*e.g.* SMILES) directly into *ab initio* quality stacking interaction energies, thereby providing a means of rapidly ranking large sets of heterocycles with regard to their stacking ability.

Introduction

Stacking interactions (*i.e.* roughly parallel face-to-face interactions between planar π-systems)[70, 137-138] play vital roles in many chemical and biological systems, and the ability to predict and rationally tune the strength of these interactions is important in everything from improving the performance of asymmetric catalysts[97, 139] to drug design.[58, 60-61] Various classes of stacking interactions are particularly in protein-ligand interactions, and harnessing these interactions can provide a powerful means of tuning inhibitor binding.[27] Chief among these are stacking interactions of heterocycles with the aromatic amino acid side chains Phe, Tyr, and Trp (see Figure 5.1).[30, 58, 60-61]

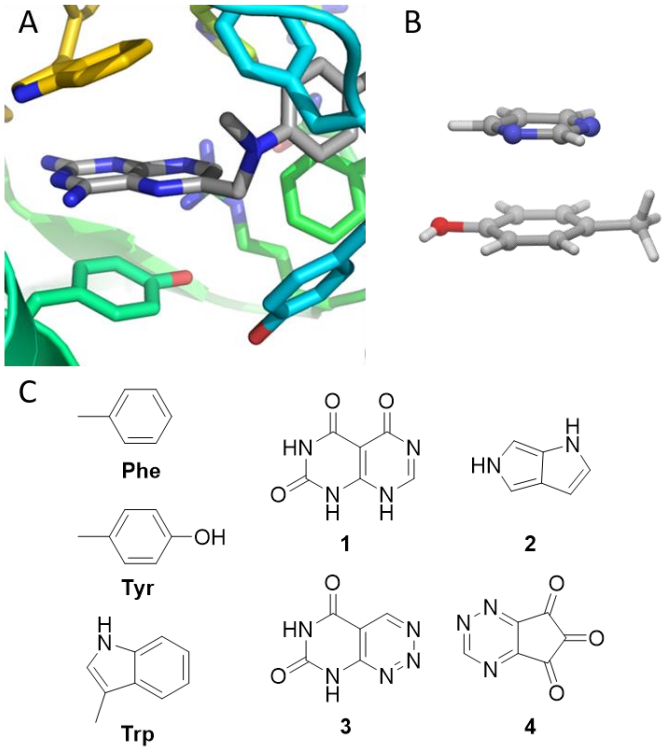


Figure 5.1. A) Example of stacking interactions of pteridine fragment of methotrexate stacking with Tyr and Trp residues in the human folate receptor (PDB: 4KNO);[148] B) model stacked dimer of pyrimidine with Tyr from Ref. [217]; and C) selected heterocycles discussed in the text as well as the model amino acid side chains Phe, Tyr, and Trp.

Tools exist for enumerating regioisomeric structures of heterocyclics in order to build greater structural complexity during exploratory chemistry and lead optimization.[218] However, there is currently no means of rapidly predicting how strongly such heterocyclic fragments will stack with binding site aromatic residues. While quantum mechanical (QM) methods can provide highly-accurate gas-phase interaction energies for such interactions, this typically requires expensive *ab initio* methods paired with large basis sets or dispersion-corrected density functional theory (DFT) methods.[131, 219-220] Moreover, finding the maximum possible stacking interaction between a pair of π-systems requires the systematic exploration of local stacked energy minima in order to

identify the global minimum energy stacked structure.[30, 172-173, 217, 221] This is a time consuming and computationally costly process.

Recently, we introduced a series of heterocycle descriptors based on the electrostatic potential (ESP) computed in the plane 3.25 Å from the heterocycle.[173] These descriptors have proved useful in the development of quantitative predictive models of the strength of stacking interactions between drug-like heterocycles and protein amide backbones,[172] Asp-Arg salt-bridges,[173] and the aromatic amino acid side chains Phe, Tyr, and Trp.[217] For example, equation 5.1 provides accurate interaction energies (RMSE = 0.8 kcal mol^{-1}) for stacked dimers of drug-like heterocycles with aromatic amino acids based on the heterocycle descriptors ESP_{mean} and ESP_{max} as well as the heavy atom counts of the heterocycle and amino acid side chain (N_{HA}^{Het} and N_{HA}^{AA}, respectively).[217,222] ESP_{mean} is the mean value of the ESP in the plane 3.25 Å from the heterocycle whereas ESP_{max} is the maximum ESP value in this plane. Figure 5.2A shows stacking interaction energies for a representative set of drug-like heterocycles predicted using equation 5.1 plotted against robust DLPNO-CCSD(T)/cc-pVQZ interaction energies.[87-90, 112, 217] The use of equation 5.1, which depends on ESP_{mean} and ESP_{max} values computed at the ωB97X-D/def2-TZVP level of theory,[110-111] provides direct access to the maximum possible stacking ability of a given heterocycle without the need to explore potential dimer geometries. The result is a ~30,000 fold reduction in computational cost, compared to DFT-based geometry optimizations of stacked dimers followed by high-accuracy *ab initio* interaction energies. More qualitatively, equation 5.1 indicates that heterocycles with increased ESP_{max} and ESP_{mean} values will exhibit stronger stacking interactions.

$$\Delta E_{pred} = N_{HA}^{AA}\left(-0.02ESP_{max} - 0.02ESP_{mean} - 0.10N_{HA}^{Het}\right) - 1.53 \quad (5.1)$$

Despite the potential utility of equation 5.1, the use of DFT-based descriptors prohibits the evaluation of stacking interactions for large libraries of heterocycles. Herein, we show that ESP_{mean} and ESP_{max}, and the corresponding stacking interaction energies, can be accurately evaluated directly from the atom connectivity for monocyclic and bicyclic heterocycles, obviating the need for quantum chemical computations.[223]

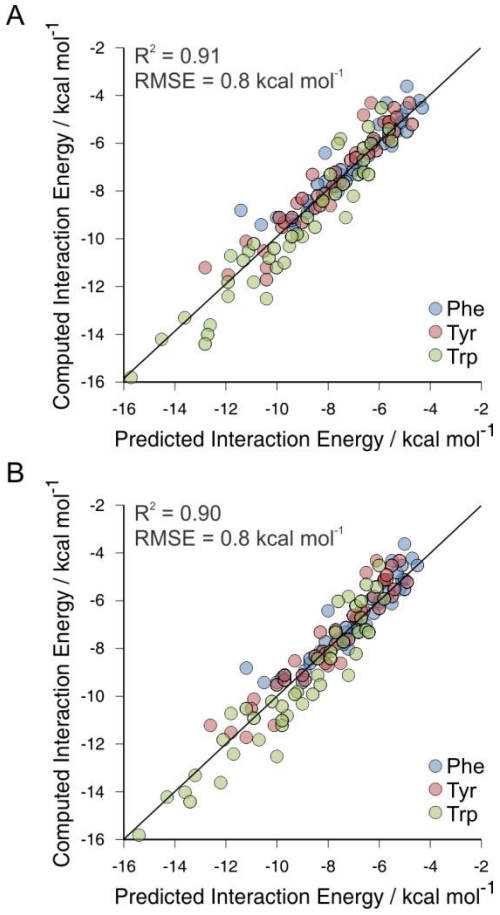


Figure 5.2. Stacking interactions predicted based on equation 5.1 versus DLPNO-CCSD(T)/cc-pVQZ interaction energies for stacking interactions of 47 drug-like heterocycles with model Phe, Tyr, and Trp side chains using A) DFT-computed ESP_{mean} and ESP_{max} values[217] and B) ESP_{mean} and ESP_{max} values evaluated using equations 5.2 and 5.3.

Results and Discussion

Below, we develop predictive models of the heterocycle descriptors ESP_{mean} and ESP_{max} based on the atom connectivity for a given heterocycle. These models are parameterized based on a diverse set of 1854 heterocycles from Pitt *et al.*[224] (see Computational Methods for more details). The DFT-computed ESP_{mean} and ESP_{max} values are presented in Figure 5.3, grouped by whether or not the heterocycle was known or unknown as of publication of Ref. [224].

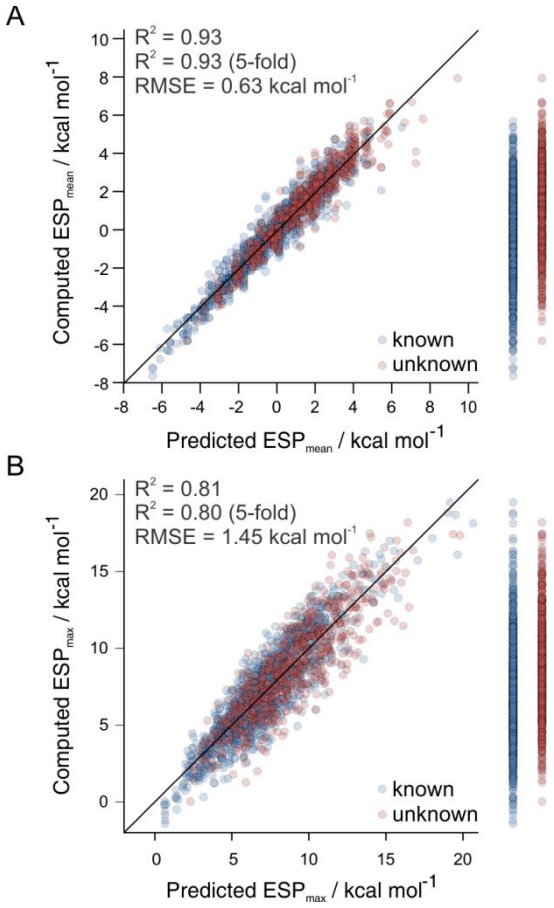


Figure 5.3. DFT computed values versus values from equations 2 and 3 for A. ESP_{mean} and B ESP_{max}. The strip plots on the right show the same data grouped by whether the heterocycles were known or unknown as of publication of Ref. [224].

Predicted stacking interaction energies of all 1854 heterocycles with Phe, Tyr, and Trp, based on equation 5.1 using the DFT-computed descriptors, are plotted in Figure 5.4.

These computed descriptors and the resulting stacking interactions from equation 5.1 enable comparisons of electrostatic characteristics and potential stacking interactions of known versus unknown heterocycles. With regard to ESP_{max}, the known heterocycles actually span a larger range of values than the unknown heterocycles, from –1.4 kcal mol^{-1} for benzene to +19.5 kcal mol^{-1} for **1** (see Figure 5.1C). Values of ESP_{mean}, on the other hand, are systematically higher for the unknown compared to the previously synthesized heterocycles and there are a number of synthetically tractable yet unknown heterocycles that exhibit very large ESP_{mean} values. For instance, ESP_{mean} values for known heterocycles range from –7.6 kcal mol^{-1} for **2** to +5.7 kcal mol^{-1} for **3**, while ESP_{mean} for the unknown heterocycles are as large as +8.0 kcal mol^{-1}. However, these heterocycles with very large ESP_{mean} values tend to include extreme degrees of functionalization with clustered carbonyl groups (*e.g.* **4**), so are likely not attractive synthetic targets. Moreover, while unknown heterocycles are among those predicted to exhibit very strong stacking interactions (see Figure 5.4), none are stronger than known heterocycles. For instance, the strongest predicted stacking interactions occur for **1**, which has previously been made and has a predicted stacking interaction with Trp of –18.4 kcal mol^{-1}. Of course, the range of gas-phase interaction energies displayed in Figure 5.4 is larger than what will be possible in the dielectric environment of a protein; however, previous solution-phase computations[217] suggest that these trends will generally hold up in protein-like dielectric environments.

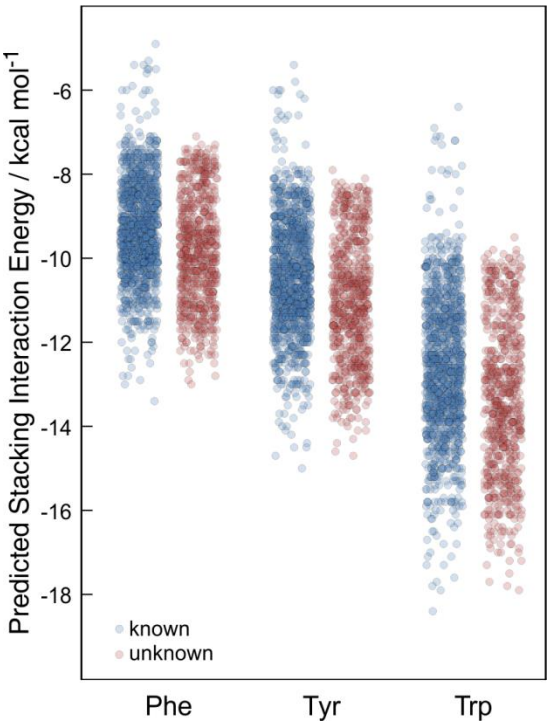


Figure 5.4. Predicted stacking interactions of 1854 known and unknown[224] heterocycles with Phe, Tyr, and Trp based on equation 1 and QM-computed ESP_{mean} and ESP_{max} values.

A. Prediction of ESP_{mean} and ESP_{max} from Atom Connectivity

Previously, we showed[217] qualitatively that ESP_{mean} reflects the number of O, S, amino N (-N=), imino N (-NH-), and carbonyl groups within a heterocycle, while ESP_{max} depends on both the number and relative position of these heteroatoms. We sought to develop a quantitative relationship between these descriptors and the atom connectivity of a given heterocycle in order to provide a rapid means of their evaluation.[225] We first define seven atom types covering those present in drug-like heterocycles (see Table 5.1; throughout, carbonyl groups are considered a single 'heteroatom' located at the position of the carbonyl carbon). We then fit a_0 as well as parameters for each of the seven atom types (a_i) to the ωB97X-D/def2-TZVP computed ESP_{mean} values for the full set of 1854 heterocycles according to equation 5.2. The optimized parameters are listed in Table 5.1;

ESP_{mean} values predicted from equation 5.2 are plotted versus the DFT computed values in Figure 5.3A. Overall, this fit is highly accurate (R^2 = 0.93, RMSE = 0.63 kcal mol^{-1} and robust to 5-fold cross validation (see Figure 5.3A).

$$ESP_{mean}^{pred} = a_0 + \sum_{i}^{atoms} a_i \qquad (5.2)$$

Errors in ESP_{mean} values predicted from equation 5.2, compared to the DFT-computed values, are as large as 3.7 kcal mol^{-1}. These errors grow with the size of ESP_{mean} (see Appendix D, Figure D-1 and Figure D-2), which arises because the heterocycles with large positive ESP_{mean} values tend to have a large proportion of heteroatoms. Equation 5.2 assumes additivity of effects for each heteroatom, but as the number of heteroatoms increases this additivity assumption slowly begins to break down. The application of equation 5.2 to predict ESP_{mean} for two representative heterocycles is demonstrated in Figure 5.5.

Table 5.1. Optimized parameters for each heteroatom type for equations 5.2 and 5.3.

Atom types	a_i	b_i	c_i
C (aryl)	–0.22	-	-
C=O	2.95	3.03	1.56
N: (imino)	1.59	1.62	1.43
O	0.78	0.75	0.56
S	0.09	0.00	0.42
NH (amino)	–0.82	1.76	-1.60
N (ring fusion)	–0.05	1.16	-0.66
	a_0 = –3.53	b_0 = 0.67	

Predicting ESP_{max} is more complex, and requires both the number of each heteroatom type (aryl carbons contribute a constant amount to ESP_{max}, regardless of their number) and the relative position of pairs of heteroatoms. We fit b_0 and coefficients for each of the six heteroatom types (b_i and c_i) according to equation 5.3. The constants α_{ij}

and R_{ij} are the relative angle and distance between radial vectors located at the positions of the two heteroatoms derived from idealized hexagonal and pentagonal rings of edge length 1.0 (see Figure 5.5). Fitted parameters are listed in Table 1; values of ESP_{max} predicted from equation 5.3 are plotted against the DFT computed values in Figure 5.3B. While the errors in ESP_{max} are slightly larger than those for ESP_{mean}, this fit is still accurate (R^2 = 0.81 and RMSE = 1.45 kcal mol^{-1}) and robust to 5-fold cross-validation (see Figure 5.3B). Errors in predicted ESP_{max} can be as large as 4.7 kcal mol^{-1}. Amongst the heterocycles with the largest errors are bicyclic systems with highly asymmetric substitution patterns (*i.e.* all heteroatoms clustered on one ring). The use of equation 5.3 to evaluate ESP_{max} for two representative heterocycles is shown in Figure 5.5.

$$ESP_{max}^{pred} = b_0 + \sum_{i}^{hetatm} b_i + \sum_{i<j}^{hetatm} \frac{cos(\alpha_{ij})}{R_{ij}} c_i c_j \qquad (5.3)$$

B. Converting SMILES to Stacking Interaction Energies

The heterocycle descriptors ESP_{mean} and ESP_{max} can be accurately predicted using equations 2 and 3 based on the number of heteroatom types and their relative positions, which can be readily derived from the connectivity information contained in SMILES representations. The resulting descriptors (along with the heavy atom count) can then be used to predict the ability of a given heterocycle to engage in stacking interactions with Phe, Tyr, and Trp based on equation 5.1. This provides a means of rapidly converting SMILES representations directly into stacking interaction energies without resorting to any quantum chemical computations.

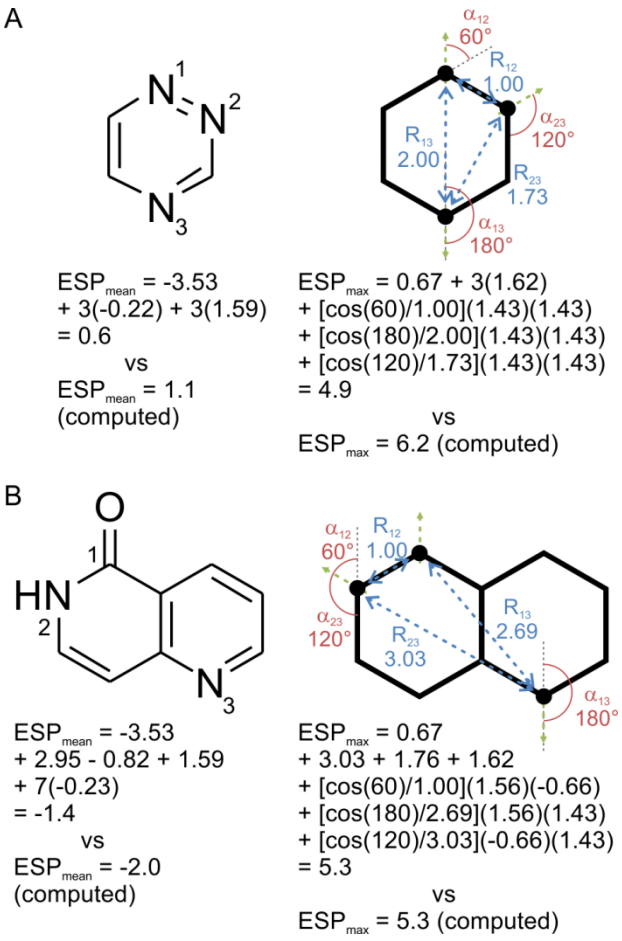


Figure 5.5. Application of equations 5.2 and 5.3 to predict ESP_{mean} and ESP_{max} (in kcal mol^{-1}) for A) 1,2,4-triazine and B) 1,6-naphthyridin-5(6*H*)-one. DFT-computed values are also provided for comparison.

To demonstrate this, Figure 5.2B shows stacking interactions evaluated directly from heterocycle SMILES representations via equations 1-3 plotted against the high-level DLPNO-CCSD(T)/cc-pVQZ data from Ref. [217]. These stacking interactions are captured with no loss in accuracy (RMSE = 0.8 kcal mol^{-1} regardless of how the descriptors are evaluated) but a substantial reduction in computational cost. More precisely, while the use of DFT-based descriptors provided a ~30,000 fold reduction in computational cost, compared to explicit optimizations of stacked dimers,[217] the evaluation of stacking interactions directly from SMILES representations provides an additional ~20,000 fold

reduction. This overall 10^9 speedup means that the prediction of stacking interaction energies for the 1854 heterocycles considered with Phe, Tyr, and Trp (*i.e.* Figure 5.4) requires less than a minute of CPU time on a decade old server.

Summary and Concluding Remarks

Predicting the potential strength of stacking interactions between heterocyclic systems is vital for a number of fields, including drug design. We previously showed that heterocycle descriptors based on the ESP evaluated 3.25 Å from the molecular plane provide a powerful means of predicting stacking interactions of drug-like heterocycles with aromatic amino acid side chains.[217] This provided access to reliable stacking interaction energies based solely on computed electrostatic potentials of the isolated heterocycles, without the need for DFT-based geometry optimizations and high-level *ab initio* interaction energies. However, the requirement to compute the ESP using quantum chemical methods prohibited the rapid evaluation of these descriptors and the resulting stacking interactions for large sets of heterocycles. Here, we showed that these ESP-based descriptors follow systematic trends based on two features of electrostatic potentials: the effect replacing a CH group with a heteroatom can be approximated by a local radial dipole located at the position of the heteroatom[226] and the impacts of multiple heteroatoms are additive. These simple features motivated the development of quantitative models that enable the evaluation of the heterocycle descriptors ESP_{mean} and ESP_{max} based solely on atom connectivity. That is, one can convert simple molecular representations (*e.g.* SMILES) directly into high-level *ab initio* quality stacking interactions with minimal loss in accuracy. This provides a means of rapidly evaluating heterocycle stacking interaction energies, opening the door for the rapid ranking of large

sets of drug-like heterocycles with regard to their potential stacking interactions with aromatic binding site residues.

Computational Methods

We selected 1854 heterocycles (51 monocycles and 1803 bicycles) from the VEHICLe database of Pitt *et al*.[227] These heterocycles all have P > 0.95, a fully aromatic canonical SMILES string, at least two (for monocycles) or three (for bicycles) unsubstituted positions, fewer than four carbonyl groups, and fewer than six imino nitrogen atoms. This provides a diverse set of heterocycles comprising 1174 known (previously synthesized) heterocycles and 680 heterocycles that had not been synthesized as of 2009 but should be synthetically viable.[227] Initial three dimensional structures of these heterocycles were generated from the SMILES representations in the VEHICLe database[227] using OpenBabel.[224] These initial structures were then optimized at the ωB97X-D/def2-TZVP level of theory[110-111] using Gaussian09.[94] All heterocycles were considered to be planar (*e.g.* no pyramidilization of amino nitrogens, *etc*.). Values for ESP_{mean} and ESP_{max} were evaluated for each heterocycle based on the electrostatic potential computed on a grid of points (0.1 Å spacing) located 3.25 Å from the heterocycle at the ωB97X-D/def2-TZVP level of theory, as done previously.[173]

In order to identify all heteroatom types and heteroatom pairs, canonicalized SMILES strings were converted to connectivity tables using OpenBabel.[224] Atom types and pairs were then identified using a Perl script based on this connectivity (a web-based interface to this Perl script is available).[228] The angles (α_{ij}) and distances (R_{ij}) in equation 3 are found in Appendix D, Table D-1. The fits of equations 5.2 and 5.3 were subjected to 5-fold cross-validation. To confirm that equations 5.2 and 5.3 were not over-fit, we

generated random sets of descriptors according to a gamma distribution with $k = 0.7$ and $\theta = 1$ for the heteroatom counts and $k = 0.15$, $\theta = 1$ for the pair counts. This resulted in a distribution of digits that roughly matched the distribution found in the initial data set. We refit equations 5.2 and 5.3 to three such random data sets. The average RMSE for equation 5.2 was 2.3 kcal mol^{-1} and the average R^2 was 0.00. Using the same random data sets to fit equation 5.3, the average RMSE was 3.3 kcal mol^{-1} and the average R^2 = 0.01.

CHAPTER 6

ENANTIOSELECTIVE CATALYST DESIGN THROUGH QUANTUM MECHANICAL VIRTUAL SCREENING [e]

[e] Bootsma, A. N., Nguyen, T. N., Guan, Y., May, J. A. and S. E. Wheeler. In preparation for submission to ACS Catalysis

Abstract

Despite the potential of computational quantum chemistry to drive the design of asymmetric catalysts, examples of prospective computational catalyst design are rare. We show that automated quantum chemical screening of a virtual library of catalysts is a practical means of designing new catalysts and demonstrate its use in the design of catalysts for the asymmetric addition of vinyl boronic acids to heterocycle-appended enones. We use the computational toolkit AARON to predict activities and enantioselectivities for a large library of potential catalysts for this reaction, identifying one predicted to offer improved performance compared to previously-known catalysts. We then use these data to derive a multivariate linear model of catalytic activity for this reaction that informs the development of a stereoelectronic model of selectivity. Finally, we experimentally test the most promising new catalysts and confirm their predicted performance, closing the loop on an iterative design process driven by quantum chemistry.

Introduction

The efficient synthesis of complex chiral molecules remains at the forefront of challenges in synthetic organic chemistry. Computational quantum chemistry has long held the promise to accelerate the design of more effective chiral catalysts,[45-46, 180] and computational catalyst design has been deemed a 'holy grail' in the field[46, 180] since it would reduce the need to synthesize and test catalysts that will ultimately prove ineffective. The underlying idea is that catalytic activities and selectivities can be predicted for previously untested catalysts by computing the free energies of stereocontrolling and rate-limiting TS structures using density functional theory (DFT). If such predictions were made for a virtual library of potential catalysts, akin to what is

done in computational drug design,[181] one could prioritize the synthesis and testing of only the most active and selective catalyst designs. Unfortunately, reliable quantum chemical predictions of activity and selectivity for a given reaction require the geometry optimization of all accessible TS structures, which can number in the hundreds for even modestly flexible molecular systems.[182-183] The result is that making DFT-based predictions for more than a handful of potential catalysts using conventional quantum chemical tools is more arduous than simply performing the required experiments, and instances of truly prospective computational catalyst design are rare.[45-46, 180, 184-190]

Previous examples of DFT-based catalyst design[184-190] typically proceeded by first studying known catalysts in order to develop a conceptual model of the factors that control activity or selectivity and then computationally 'testing' a small number of potential new designs chosen based on this conceptual model. For instance, Ito and co-workers[188] recently designed highly effective chiral bisphosphine ligands for enantioselective Markovnikov hydroborations of terminal aliphatic alkenes using a combination of DFT and experiment. They considered five potential designs in the development of a steric model of a quinoxaline-based ligand. Unfortunately, many reactions will not be amenable to the development of a robust model of activity and selectivity based on such a small number of examples and the routine identification of superior catalysts will typically require consideration of a much larger number of potential designs.

Norrby and co-workers[191] have developed an alternative approach to computational catalyst design based on the screening of virtual libraries of potential catalysts using QM-derived molecular mechanics force fields (Q2MM).[192] Such force-

fields can provide accuracy rivaling DFT but at a drastically reduced computational cost. Norrby *et al*.[191] recently demonstrated the power of their CatVS tool[191] to computationally screen virtual libraries of a dozen substrates or ligands, delivering reliable predictions of selectivity for several reactions. However, such MM-based screening approaches are most well-suited for systems for which there is a single stereodetermining TS across all combinations of catalyst and substrate[191] and when one is only interested in selectivity but not catalytic activity. Moreover, application of CatVS requires the development of an MM force field for each new reaction type considered,[192] which represents a major bottleneck in the widespread application of this approach. An analogous, fully QM-based virtual screening protocol would not be subject to such limitations.

We recently introduced an open-source computational toolkit (AARON: An Automated Reaction Optimizer for New catalysts)[193] that automates the DFT-based optimization of TS structures for homogeneous catalytic reactions based on previously-computed TS structures for a representative model catalyst. This automation removes the burden of manually performing geometry optimizations of hundreds of TS structures, opening the door for rapid QM-based predictions of both activities and selectivities of virtual libraries of dozens of catalysts for a given reaction.[52, 97] Below, we describe a new catalyst design approach combining such QM-based virtual screening with multivariate linear regression (MLR) methods[194] to deliver both improved catalysts and an *understanding* of why these catalysts outperform others. The utility of this approach is demonstrated by designing new catalysts for the asymmetric addition of vinyl boronic acids to heterocycle-appended enones, with experimental verification.

Results

May *et al*.[195-196] developed a method for the enantioselective addition of vinyl boronic acids to heterocycle-appended enones catalyzed by 3,3′-substituted BINOLs (Figure 6.1A), providing access to a wide range of α-chiral heterocycles. The observed activity and selectivity of this reaction is sensitive to the nature of the 3,3′-substituents (R) on the catalyst (Table 6.1), with the most effective catalysts featuring fluorinated aryl groups at these positions (Figure 1B). While **Cat5** delivered the desired stereoisomer in good yield, there remains a need for more active catalysts for this reaction to avoid degradation of the starting materials. Goodman *et al*.[197-199] had previously studied related additions of alkenyl boronate esters to β-arylenones,[200-201] attributing the activity of **Cat2** in part to the ability of the 3,3′-iodo substituents to stabilize the build-up of a formal negative charge on the boron during the reaction. The selectivity was explained based on a simple steric model in which one of the 3,3′-substituents blocks one face of the enone. Based on this previous work,[197-199] May *et al*.[195] was able to explain the trend in reactivity seen in Table 6.1 in terms of the increasing electron-withdrawing ability of the 3,3′-substituents. However, the applicability of Goodman's steric model to the selectivity of these reactions is less obvious, suggesting that there is more at play than simple steric effects.

By analogy with Goodman's work,[197-199] the first step in the mechanism for reaction 1 (see Figure 6.1C) is the condensation of the catalyst with styrenyl boronic acid **2** to give a chiral boronate ester **4**. This boronate ester then forms a pre-reaction complex **5** with the heteroaryl enone **1**, which goes through the rate-limiting and stereocontrolling C–C bond forming TS to give boron enolate **6**. Protonolysis of **6** releases the catalyst and

the desired chiral product **3**. The restricted geometry of the cyclic boronate ester **4** leads to poor electron donation from the BINOL oxygens.[197-199] This, combined with the electron-withdrawing character of the 3,3′-substituents, enhances the Lewis acidity of the boron and activates it towards the addition.

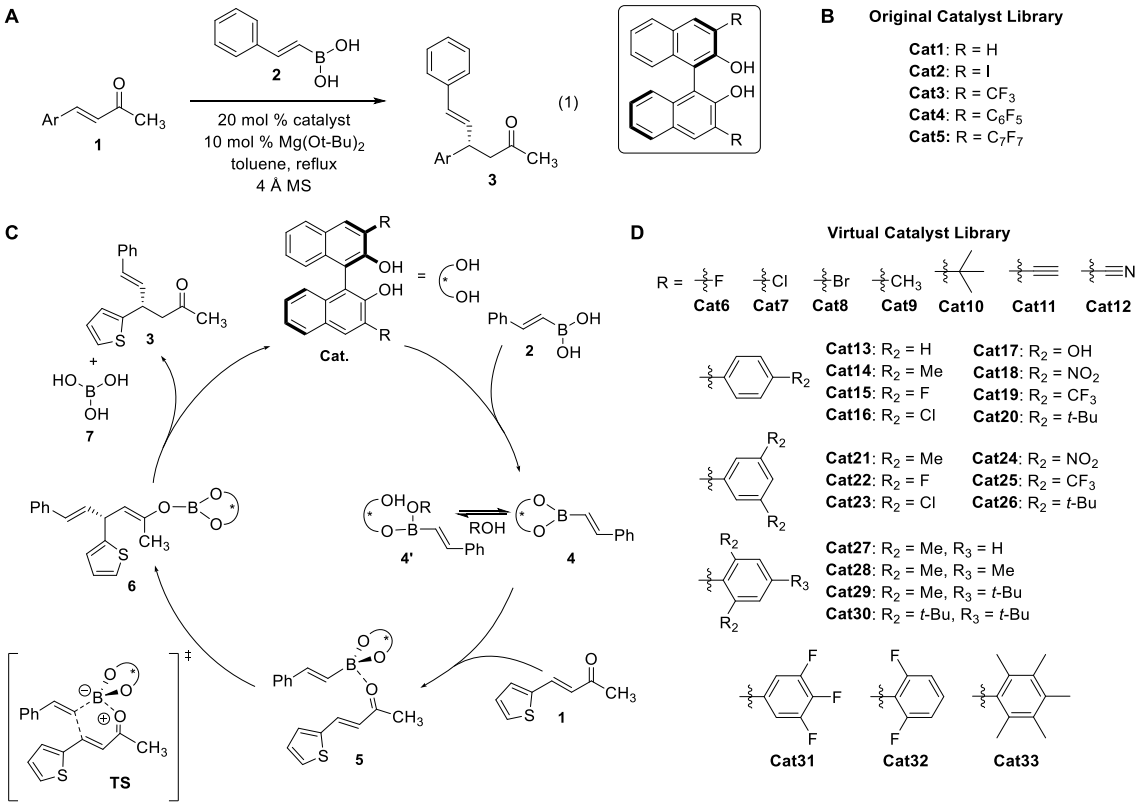


Figure 6.1. A. BINOL catalyzed enantioselective addition of vinyl boronic acids to heterocycle-appended enones from May *et al*.[195] B. Original library of catalysts considered experimentally by May *et al*.[195]; C. Catalytic cycle for reaction 1 (Ar = 2-thiophene); D. Virtual catalyst library screened for reaction 1.

The comparable single ligand exchange boronate ester **4′** is lower in energy, but the cyclic ester is more reactive so this pathway is preferred. The geometry of the transition state is restricted by the conjugated, planar nature of the substrate and so the reaction occurs through a 'sofa-like' TS.[197-199]

Table 6.1. Experimental er values and yields for reaction 1 (Ar = 2-thiophene) along with computed er values, the difference in free energy barriers between stereocontrolling transition states ($\Delta\Delta G^{\ddagger}$, in kcal mol^{-1}), and the overall barrier for TS-Re ($\Delta G^{\ddagger}$, in kcal mol^{-1}).

	Exp.		Theory		
Catalyst	*er*[a]	Yield[a]	*er*[b]	$\Delta\Delta G^{\ddagger}$	$\Delta G^{\ddagger}$
Cat1	73:27	15% (6 h)	80:19	2.1	37.5
Cat2	96:4	40% (6 h)	98:2	3.3	30.7
Cat3	95:5	44% (6 h)	97:3	3.1	32.2
Cat4	97:3	58% (6 h)	>99:1	5.2	27.6
Cat5	96:4	87% (4 h)	>99:1	5.5	25.8
***t*-BuO$^-$**				0.0	38.2

[a] Experimental data from Ref. [195]. [b] Theoretical *er* values account for a Boltzmann weighting over all accessible TS structures as well as the *t*-BuO^- catalyzed background reaction.

Our computations show that each intermediate leading to the C–C bond forming TS is higher in free energy than the separated reactants, so the overall activity is determined by the free energy difference between the TS structure and separated reactants ($\Delta G^{\ddagger}$). This is corroborated by experimental Hammett plot analysis from May *et al.*[202] that supports C–C bond forming as rate-determining. Boronate ester **4** was also confirmed as the primary reacting species both through computation of the potentially competitive pathway using a monodentate ester and through experimental studies with a monoalkylated analog of **Cat4**, which reacted slowly and non-selectively.

To ensure accurate free energy barriers, all accessible transition state confirmations need to be considered. With two rotatable groups present in the substrate, there are a minimum of eight distinct TS geometries for each catalyst—four leading to the *Re* adduct and four leading to *Si*. All eight of these TS structures (plus all catalyst rotamers) were automatically located for **Cat1**-**Cat5** using AARON[193] based on previously computed TS structures for a representative model catalyst. As seen in Table 6.1, the difference in free energy between the lowest-lying *Re* and *Si* TS structures

($\Delta\Delta G^{\ddagger}$) is substantial even for the least selective catalyst (**Cat1**). This is in conflict with Goodman's steric model,[197-199] because **Cat1** features hydrogens at the 3,3′ positions, and again suggests that other factors underlie the selectivity. However, examination of the stereocontrolling TS structures does not reveal any obvious source of these free energy differences, particularly for **Cat1**.

Selectivities predicted based solely on $\Delta\Delta G^{\ddagger}$ greatly overestimate experimental *er* values (see Appendix E, Table E-1) and using a Boltzmann weighting over all conformations offers no improvement on these results. These predictions, however, are based on the assumption that only transition states involving the BINOL catalyst contribute to product formation. Based on the low yields that correspond to the less selective catalysts (see Table 6.1), it was postulated that an inability to outcompete the racemic background reaction contributes to the lower observed enantiomeric ratios. Experimentally, increased concentration of Mg(*t*-BuO)$_2$ leads to higher yield and lower selectivity, indicating that *t*-BuO$^-$ plays a role in a non-selective pathway. We located a TS structure for the *t*-BuO$^-$ catalyzed background reaction (see Appendix E, Figure E-1), and the associated barrier of 38.4 kcal/mol is low enough to affect the observed selectivity for some of the catalysts. For instance, the free energy barrier for the *t*-BuO$^-$ catalyzed reaction lies only 0.3 kcal mol^{-1} higher in free energy than TS-*Re* for **Cat1**. Thus, the selectivity of **Cat1** is determined primarily by this free energy difference, not $\Delta\Delta G^{\ddagger}$. In other words, **Cat1** is selective with regard to *Re*- vs *Si*-facial attack but is too inactive for this selectivity to come to fruition. Once we account for the *t*-BuO$^-$ catalyzed background reaction, predicted *er* values are in good agreement with experiment (see Table 6.1), although they overestimate the selectivity of the highly-selective catalysts.

Given the relatively large $\Delta\Delta G^{\ddagger}$ values for all of the catalysts, the catalyst activity (*i.e.* its ability to out-compete the *t*-BuO^- catalyzed reaction) is the primary selectivity determining factor. May *et al.*[195] previously proposed that the increasing activity going from **Cat1** to **Cat5** was due to increasing Lewis acidity of the boron upon addition of more strongly electron withdrawing substituents.[197-199] To evaluate this quantitatively, we used fluoride affinities of the boronate esters **4** as a measure of Lewis acidity.[203] The activation free energies for reaction 1 catalyzed by **Cat1**-**Cat5** are plotted against the fluoride affinities of the corresponding boronate esters in Figure 6.2A (red dots). There is a strong linear correlation ($R^2 = 0.98$), which seems to confirm the proposed Lewis acidity based model for this reaction.[195, 197-199]

To probe this relationship further, and potentially identify more effective catalysts for reaction 1, we screened a library of 28 substituted BINOLs for which experimental data for reaction 1 were unavailable (Figure 6.1D). For each catalyst in this virtual library, we used AARON[193] to automatically compute all low-lying conformations of the stereocontrolling TS structures, including all conformations of rotatable substituents (*e.g.* *t*-Bu, CF_3, *etc*.). In total, AARON located 391 unique TS structures using eight manually computed TS structures as templates. Predicted *er* values range from 77:23 to >99:1 and $\Delta G^{\ddagger}$ from 23.4 to 37.5 kcal mol^{-1} (see Table 6.2). Several examples highlight the need for both a highly active catalyst and a large $\Delta\Delta G^{\ddagger}$ value to achieve excellent selectivity, which are both predicted using this QM-based screening. For instance, for **Cat10** we predict a significant free energy difference between TS-*Re* and TS-*Si* but low activity. The result is a predicted *er* of 91:9. On the other hand, for **Cat26** we predict a low

reaction barrier (26.6 kcal mol^{-1}) but a modest $\Delta\Delta G^{\ddagger}$, resulting in a predicted *er* of only 84:16.

Table 6.2. DFT computed ($\Delta G^{\ddagger}_{calc}$) and predicted ($\Delta G^{\ddagger}_{pred}$, from equation 6.1) free energy barrier heights and DFT computed $\Delta\Delta G^{\ddagger}$ values in kcal mol^{-1}, and theoretical *er* values for **Cat6-Cat33** as well as experimental data for selected catalysts.

Cat.	Theor *er*	$\Delta\Delta G^{\ddagger}$	$\Delta G^{\ddagger}_{calc}$	$\Delta G^{\ddagger}_{pred}$	Exp % Yield	Exp *er*
Cat6	88.7:11.3	1.9	33.2	31.3		
Cat7	93.7:6.3	2.4	32.0	31.5		
Cat8	96.4:3.6	2.6	31.9	31.6		
Cat9	97.6:2.4	4.2	35.6	34.8		
Cat10	90.9:9.1	6.6	37.1	35.3		
Cat11	93.1:6.9	2.2	32.8	32.4		
Cat12	76.9:23.1	0.9	29.1	28.1		
Cat13	98.8:1.2	2.9	29.2	28.8		
Cat14	98.5:1.5	2.6	27.2	29.3		
Cat15	99.0:1.0	3.0	28.7	28.9		
Cat16	>99:1	3.3	27.9	28.2		
Cat17	98.5:1.5	2.8	29.6	29.3		
Cat18	>99:1	4.5	26.5	26.4	79[a]	92:8
Cat19	>99:1	4.3	27.4	27.5		
Cat20	>99:1	4.4	26.5	26.2	51[b]	79:21
Cat21	98.6:1.4	2.9	28.1	28.1		
Cat22	>99:1	3.6	27.6	27.3		
Cat23	>99:1	3.6	26.2	26.7	84[c]	95.5:4.5
Cat24	>99:1	6.1	23.4	24.2	93[d]	98.5:1.5
Cat25	>99:1	3.8	25.1	24.4	82[c]	96:4
Cat26	84.3:15.7	1.4	25.1	26.6		
Cat27	>99:1	4.5	30.9	29.4		
Cat28	>99:1	5.6	29.2	28.8		
Cat29	>99:1	4.6	28.8	29.0		
Cat30	98.9:1.1	3.2	32.2	32.3		
Cat31	>99:1	3.7	25.7	26.9		
Cat32	>99:1	4.5	29.9	31.1		
Cat33	>99:1	4.0	30.4	29.2		

Reaction times: [a]10 h; [b]24 h; [c]6 h; [d]1.5 h.

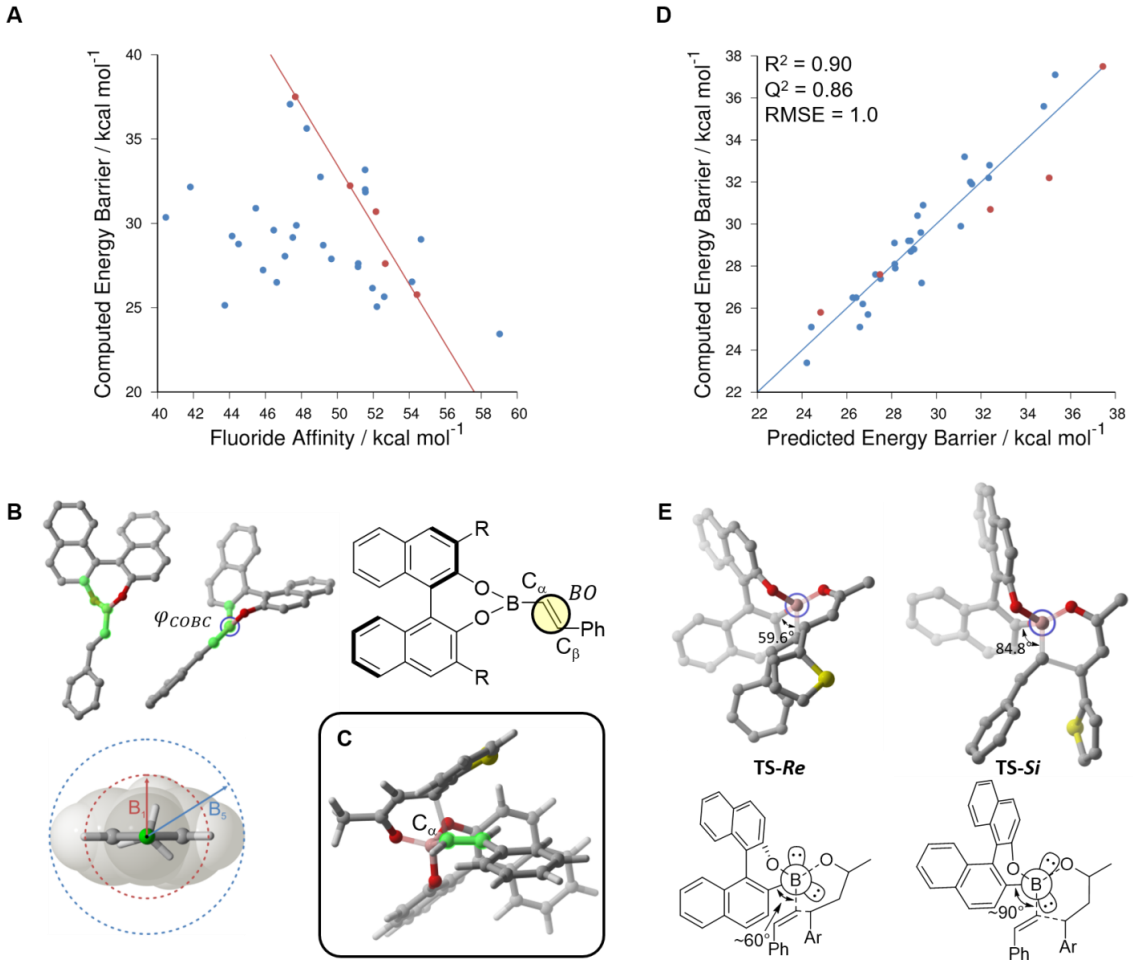


Figure 6.2. A. Computed free energy barriers versus the fluoride affinities of the corresponding boronate esters for **Cat1**-**Cat5** (red) and **Cat6**-**Cat33** (blue). The red line is the best-fit line through the data for **Cat1**-**Cat5** only (R^2 = 0.98); B. Definition of the parameters φ_{COBC}, B_1, and BO used in the fit in equation 1; C. TS-*Re* structure for **Cat1** showing the distorted tetrahedral geometry around C_α (the C_α–C_β bond is highlighted). D. Plot of computed free energy barriers for **Cat1**-**Cat5** (red) and **Cat6**-**Cat33** (blue) versus predicted free energy barriers from equation 6.1; E. Stereochemical model of the selectivity of BINOL based catalysts for reaction 1 (viewed down one of the B–O bonds) in which the BINOL oxygen lone pair is in an *anti*-periplanar alignment with the breaking B–C bond in the favored TS-*Re* but not in TS-*Si*.

The free energy barriers for this expanded library of catalysts are plotted versus the fluoride affinities of the boronate esters in Figure 6.2A (blue dots). While the barrier heights for **Cat1**-**Cat5** showed an excellent correlation with fluoride affinities, this trend breaks down for this more diverse set of catalysts. Apparently, the correlation observed

for the original five catalysts stems from the fact that these varied primarily in terms of Lewis acidity. Data from the full set of catalysts indicates that tuning the activity of BINOL catalysts for reaction 1 is more complicated than simply ramping up the Lewis acidity and we predict several catalysts with relatively modest Lewis acidity but high activity as well as systems with high Lewis acidity and relatively low activity. For instance, **Cat14** has a fluoride affinity of only 57.6 kcal mol^{-1} (compared to 66.0 kcal mol^{-1} for **Cat5**), but is predicted to have an activation free energy that is within 1.5 kcal mol^{-1} of that for **Cat5**. On the other hand, **Cat12** has a fluoride affinity of 65.9 kcal mol^{-1} yet a barrier that is 3.5 kcal mol^{-1} higher than that for **Cat5**.

To identify the factors that actually determine the activity (and therefore stereoselectivity) of BINOL catalysts for reaction 6.1, we turned to MLR methods.[194] Such methods, championed by Sigman and co-workers,[38-40, 204-207] among others,[44, 208-209] can provide a powerful means of relating structure to function in complex molecular systems. However, in contrast to previous applications of MLR,[38-40, 44, 204-209] which were based on experimental data, here we sought to develop a predictive model of catalyst activity based solely on DFT-predicted $\Delta G^{\ddagger}$ values. After considering a number of geometric, steric, and electronic descriptors of the 3,3′-substituents, the BINOLs, and the corresponding boronate esters, we derived the following three-parameter model of the free energy barrier for reaction 1,

$$\Delta G^{\ddagger}_{pred} = 0.42\varphi_{COBC} + \frac{8.3}{B_1} + 530\mathrm{BO} - 1022.8 \qquad (6.1)$$

in which φ_{COBC} is the C-O-B-C dihedral angle in the boronate ester, B_1 is the Sterimol parameter[210] of the 3,3′-substituents on the BINOL, and BO is the C_α-C_β bond order in the boronate ester (see Figure 6.2B).

Free energy barriers predicted using equation 1 are plotted against the DFT-computed data in Figure 6.2D and show a strong correlation ($R^2 = 0.90$). Moreover, this fit is robust to cross-validation ($Q^2 = 0.86$), exhibits a root mean squared error (RMSE) of 1.0 kcal mol^{-1}, and applies to the diverse set of catalysts in Figure 6.1D in addition to the original catalysts tested experimentally by May and co-workers[195] (Figure 6.1B). Thus, one can predict reliable activation free energies for reaction 1 based on three readily computed molecular descriptors (φ_{COBC}, B_1, and BO) without requiring additional DFT-based transition state optimizations. Moreover, because the aryl substituent on the enone plays no substantial role, one would expect a close analog of equation 6.1 to provide reliable predictions for other heterocycle-appended enones.

In addition to the predictive power of equation 6.1, the presence of only three simple descriptors provides key insight into the factors that determine catalytic activity. Notably, equation 6.1 does not depend on any direct measure of Lewis acidity. Instead, this fit indicates that the activity of substituted BINOLs in the catalysis of reaction 1 depends primarily on three factors:

1) the degree of overlap between the BINOL oxygen lone pair and the empty p-orbital on boron (which increases with increase in the dihedral angle φ_{COBC}), with poorer orbital overlap corresponding to lower barriers (consistent with Goodman's original description of these reactions);[197-199]
2) the ability of the 3,3′-substituents on the BINOL to create a more confined substrate binding pocket that facilitates the reaction, with larger B_1 Sterimol parameters leading to lower barriers; and

3) the ease with which C_α of the vinyl boronic acid component can be distorted into a tetrahedral geometry during the rate-limiting TS (see Figure 6.2C), as measured by the bond order of the reacting alkene in the corresponding boronate ester, with easier distortion correlating with a lower reaction barrier

Armed with this new understanding of the main factors impacting the barrier height for these reactions, we reconsidered the source of the large $\Delta\Delta G^{\ddagger}$ values favoring TS-*Re* over TS-*Si*. Given the presence of φ_{COBC} in equation 6.1, we examined the analogous dihedral angles surrounding the boron in TS-*Re* and TS-*Si*. For the disfavored TS-*Si*, the two COBC dihedral angles are tightly clustered around 90° and 180°, whereas the corresponding angles for TS-*Re* vary significantly from these values (see Appendix E, Figure E-2). For instance, for **Cat1**, one of these COBC angles is 59.6° for TS-*Re* but 84.8° for TS-*Si* (see Figure 6.2E). This is consistent with the qualitative stereoelectronic model depicted in Figure 6.1E, in which one of the BINOL oxygen lone pairs is in an *anti*-periplanar alignment with the breaking B–C bond in TS-*Re*, but not in TS-*Si*; the geometry of TS-*Si* is simply not compatible with this favorable orbital alignment. This explains why even **Cat1**, which lacks any sterically-demanding 3,3′-substituents, still exhibits a substantial difference in free energy between TS-*Re* and TS-*Si*.

The simple steroelectronic model in Figure 6.2E, combined with the insights regarding activity provided by equation 6.1, provide a complete understanding of the requirements for effective catalysts for this reaction. Such an understanding would not have been available from simply studying **Cat1**-**Cat5**, and can be used to guide the development of even more effective catalysts for this reaction. However, in this case the computational screening has already identified one catalyst (**Cat24**) predicted to provide

performance superior to **Cat5**! To verify this prediction and to test other catalysts predicted to perform well for reaction 1, we synthesized and tested **Cat18**, **Cat20**, and **Cat23**-**Cat25** (see Table 6.2). Satisfyingly, we find that all except **Cat20** exhibit activities consistent with the computational data. While the predicted *er* values overestimate the experimental data (as seen for **Cat4** and **Cat5** above), these four catalysts deliver selectivities similar to **Cat5**. Perhaps most importantly, **Cat24**, which was predicted to be more active than **Cat5**, provides 93% yield in only 1.5 hours compared to 87% yield after 4 hours for **Cat5**, with greater selectivity (98.5:1.5 vs 96:4). Thus, we have successfully identified a catalyst for reaction 1 that delivers improved activity and selectivity (with experimental verification) through the DFT-based screening of a virtual library of potential catalyst designs. Moreover, from equation 6.1 and the computed descriptors for **Cat5** and **Cat24**, we can see that the superior activity of **Cat24** arises from the smaller C-O-B-C angle and reduced bond order of the reacting alkene (BO), compared to **Cat5**.

Conclusions

Continuing advances in computational resources and algorithms have put computational quantum chemistry in a position to accelerate the design of asymmetric catalysts. However, the purely QM-based screening of virtual libraries of potential catalysts has remained just out of reach due to the time and effort required to compute all transition state structures necessary to make reliable predictions of catalytic activity and selectivity. At the same time, multivariate linear regression methods have emerged as a complementary approach[194] whereby experimental data are leveraged to developed predictive models of catalyst performance while also providing chemical insight. We

have demonstrated a catalyst design approach in which automated DFT-based screening of a virtual library of potential asymmetric catalysts is combined with MLR methods to deliver both superior asymmetric catalysts (with experimental verification) and an understanding of the origin of their enhanced performance. In the present case, screening a single virtual library delivered a better catalyst. However, in more general cases we could use the understanding delivered by the application of MLR methods to devise a refined virtual library for further screening. Such an iterative computational design process would be complementary to the CatVS approach of Norrby and co-workers,[191] but without the limitations inherent in MM-based predictions.

Looking ahead, the iterative design of catalysts by combining automated QM-based TS optimizations with MLR methods should provide a powerful means of developing catalysts with improved activity and selectivity. While we have demonstrated this approach to catalyst design for a single organocatalyzed reaction, analogous computations can be readily applied to any catalytic reaction for which the mechanism is known. In other words, this approach will be applicable to any reaction for which DFT can currently be used.[45-46, 180] In this way, we can exploit the strengths of both DFT-based computations of TS structures with applications of modern data analysis tools in order to accelerate catalyst development, bringing us one step closer to the 'holy grail' of computational catalyst design.[46]

Methods

All computations were completed using Gaussian 09 combined with the computational toolkit AARON.[193] Geometries, vibrational frequencies, and thermal free energy corrections were computed at the B97-D/def2-TZVP level of theory[111, 176-177]

using the PCM model[211] to account for solvent effects (toluene). All TS geometries exhibited a single imaginary vibrational frequency. Final reported free energies were computed at the PCM-ωB97X-D/def2-TZVP//PCM-B97-D/def2TZVP level of theory[110] with thermal free energy corrections determined using Grimme's quasi-RRHO method[212] at 384 K. All computations except for **Cat26** and **Cat30** were done with a (99,590) integration grid;[213] for **Cat26** and **Cat30** we used a smaller (72,302) grid for computational expediency. Reported $\Delta\Delta G^{\ddagger}$ values are based on the difference in free energy between the lowest-lying TS for *Si* facial attack and *Re* facial attack, *i.e.* $\Delta\Delta G^{\ddagger} = G^{\ddagger}(\text{TS-}Si) - G^{\ddagger}(\text{TS-}Re)$ for each catalyst. Theoretical *er* values were determined using a Boltzmann weighting over all TS barriers (relative to separated reactants) for *Re* or *Si* attack, including those catalyzed by *t*-BuO^{-}. Sterimol parameters[210] were computed using Sterimol v. 1.0 from Paton *et al*.,[214] while bond orders were determined using natural bond orbital (NBO) analysis.[215-216] All descriptors were computed at the same level of theory as the corresponding TS structures. Cross validation of equation 1 via L1O analysis was done using scikit-learn version 17.1.[91] Experimental selectivities for **Cat18**, **Cat20**, and **Cat23-Cat25** were determined by HPLC with chiral stationary phases and were run in triplicate.

CHAPTER 7

CONCLUSION

These five projects have demonstrated the utility of linear free energy relationships (LFERs) in a variety of contexts. They can be used both to make quantitative predictions about chemical behavior and to develop qualitative models that increase our understanding of these same phenomena. I have developed such predictors for a variety of electrostatically driven noncovalent interactions. In the case of amide stacking, I was able to apply newly developed heterocycle descriptors to improve upon a previously proposed relationship between interaction energies and molecular dipole moment, and further highlight the importance of local interactions in directing and strengthening this interaction. In the case of the salt-bridge stacking interactions, I used these heterocycle descriptors to develop both a quantitatively predictive LFER and a new model of the interaction. The combination of these two things was used to provide qualitative guidelines useful for tuning this interaction in a design context. Thirdly, I applied this methodology toward understanding arene stacking between heterocycles and aromatic amino acids. Again, I was able to develop a quantitatively predictive LFER, along with qualitative guidelines for tuning this interaction through heterocycles choice, or rapidly ranking a heterocycle set. Further, the molecular descriptors used in the development of this particular LFER are directly related to the number and location of heteroatoms in the heterocycle. This allowed me to develop simple equations that provide these descriptor values, and thus a prediction of the arene stacking interaction energy

directly from atom connectivity information, such as is provided in a SMILES string. I have also demonstrated the utility of LFERs in the area of computational catalyst design through application of quantum mechanical virtual screening for the identification of an improved catalyst followed by creation of a predictor that identified the cause of increased activity. In cases where a first round of virtual screening does not identify an improved catalyst, this methodology could be applied to enhance library selection for further screening. Ultimately these five projects have demonstrated the broad range of applicability of multiparameter LFERs both with regard to the range of chemical processes that they can be used to described, their predictive power, and their ability to provide insight and qualitative understanding rather than simply being ‘black box’ tools for quantitative predictions.

REFERENCES

1. Hammett, L. P., The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937,** *59* (1), 96-103.

2. Duan, G.; Smith, V. H.; Weaver, D. F., An ab initio and data mining study on aromatic–amide interactions. *Chem. Phys. Lett.* **1999,** *310* (3), 323-332.

3. Zimmermann, M. O.; Boeckler, F. M., Targeting the protein backbone with aryl halides: systematic comparison of halogen bonding and [small pi][three dots, centered][small pi] interactions using N-methylacetamide. *MedChemComm* **2016,** *7* (3), 500-505.

4. Kemmink, J.; van Mierlo, C. P. M.; Scheek, R. M.; Creighton, T. E., Local Structure Due to an Aromatic-Amide Interaction Observed by 1H-Nuclear Magnetic Resonance Spectroscopy in Peptides Related to the N Terminus of Bovine Pancreatic Trypsin Inhibitor. *J. Mol. Biol.* **1993,** *230* (1), 312-322.

5. Duan, G.; Smith, V. H.; Weaver, D. F., Characterization of Aromatic−Amide(Side-Chain) Interactions in Proteins through Systematic ab Initio Calculations and Data Mining Analyses. *J. Phys. Chem. A* **2000,** *104* (19), 4521-4532.

6. Duan, G.; Smith, V. H.; Weaver, D. F., A data mining and ab initio study of the interaction between the aromatic and backbone amide groups in proteins. *Int. J. Quantum Chem.* **2000,** *80* (1), 44-60.

7. Goyal, S.; Chattopadhyay, A.; Kasavajhala, K.; Priyakumar, U. D., Role of Urea–Aromatic Stacking Interactions in Stabilizing the Aromatic Residues of the Protein in Urea-Induced Denatured State. *J. Am. Chem. Soc.* **2017,** *139* (42), 14931-14946.

8. Roehrig, S.; Straub, A.; Pohlmann, J.; Lampe, T.; Pernerstorfer, J.; Schlemmer, K.-H.; Reinemer, P.; Perzborn, E., Discovery of the Novel Antithrombotic Agent 5-Chloro-N-({(5S)-2-oxo-3- [4-(3-oxomorpholin-4-yl)phenyl]-1,3-oxazolidin-5-yl}methyl)thiophene- 2-carboxamide (BAY 59-7939): An Oral, Direct Factor Xa Inhibitor. *J. Med. Chem.* **2005,** *48* (19), 5900-5908.

9. Salonen, L. M.; Holland, M. C.; Kaib, P. S. J.; Haap, W.; Benz, J.; Mary, J.-L.; Kuster, O.; Schweizer, W. B.; Banner, D. W.; Diederich, F., Molecular Recognition at the Active Site of Factor Xa: Cation–π Interactions, Stacking on Planar Peptide Surfaces, and Replacement of Structural Water. *Chem. Eur. J.* **2012,** *18* (1), 213-222.

10. Parrish, R. M.; Sitkoff, D. F.; Cheney, D. L.; Sherrill, C. D., The Surprising Importance of Peptide Bond Contacts in Drug–Protein Interactions. *Chem. Eur. J.* **2017,** *23* (33), 7887-7890.

11. Hartman, A.; Mondal, M.; Radeva, N.; Klebe, G.; Hirsch, A., Structure-Based Optimization of Inhibitors of the Aspartic Protease Endothiapepsin. *Int. J. Mol. Sci.* **2015,** *16* (8), 19184.

12. Mondal, M.; Radeva, N.; Fanlo-Virgós, H.; Otto, S.; Klebe, G.; Hirsch, A. K. H., Fragment Linking and Optimization of Inhibitors of the Aspartic Protease Endothiapepsin: Fragment-Based Drug Design Facilitated by Dynamic Combinatorial Chemistry. *Angew. Chem. Int. Ed.* **2016,** *55* (32), 9422-9426.

13. Lauber, B. S.; Hardegger, L. A.; Asraful, A. K.; Lund, B. A.; Dumele, O.; Harder, M.; Kuhn, B.; Engh, R. A.; Diederich, F., Addressing the Glycine-Rich Loop of Protein Kinases by a Multi-Facetted Interaction Network: Inhibition of PKA and a PKB Mimic. *Chem. Eur. J.* **2016,** *22* (1), 211-221.

14. Qiu, Z.; Kuhn, B.; Aebi, J.; Lin, X.; Ding, H.; Zhou, Z.; Xu, Z.; Xu, D.; Han, L.; Liu, C.; Qiu, H.; Zhang, Y.; Haap, W.; Riemer, C.; Stahl, M.; Qin, N.; Shen, H. C.; Tang, G., Discovery of Fluoromethylketone-Based Peptidomimetics as Covalent ATG4B (Autophagin-1) Inhibitors. *ACS Med. Chem. Lett.* **2016,** *7* (8), 802-806.

15. Mitchell, J. B. O.; Nandi, C. L.; McDonald, I. K.; Thornton, J. M.; Price, S. L., Amino/Aromatic Interactions in Proteins: Is the Evidence Stacked Against Hydrogen Bonding? *J. Mol. Biol.* **1994,** *239* (2), 315-331.

16. Cotesta, S.; Stahl, M., The environment of amide groups in protein–ligand complexes: H-bonds and beyond. *J. Mol. Model.* **2006,** *12* (4), 436-444.

17. Tóth, G.; Watts, C. R.; Murphy, R. F.; Lovas, S., Significance of aromatic-backbone amide interactions in protein structure. *Proteins: Struct., Funct., Bioinf.* **2001,** *43* (4), 373-381.

18. Harder, M.; Kuhn, B.; Diederich, F., Efficient Stacking on Protein Amide Fragments. *ChemMedChem* **2013,** *8* (3), 397-404.

19. Wheeler, S. E.; Bloom, J. W. G., Anion-[small pi] interactions and positive electrostatic potentials of N-heterocycles arise from the positions of the nuclei, not changes in the [small pi]-electron distribution. *Chem. Commun. (Cambridge, U. K.)* **2014,** *50* (76), 11118-11121.

20. Kaan, H. Y. K.; Major, J.; Tkocz, K.; Kozielski, F.; Rosenfeld, S. S., "Snapshots" of Ispinesib-induced Conformational Changes in the Mitotic Kinesin Eg5. *J. Biol. Chem.* **2013,** *288* (25), 18588-18598.

21. Vetter, I. R.; Arndt, A.; Kutay, U.; Görlich, D.; Wittinghofer, A., Structural View of the Ran-Importin β Interaction at 2.3 Å Resolution. *Cell* **1999,** *97* (5), 635-646.

22. Arold, S.; Franken, P.; Strub, M.-P.; Hoh, F.; Benichou, S.; Benarous, R.; Dumas, C., The crystal structure of HIV-1 Nef protein bound to the Fyn kinase SH3 domain

suggests a role for this complex in altered T cell receptor signaling. *Structure* **1997,** *5* (10), 1361-1372.

23. Crowley, P. B.; Golovin, A., Cation–π interactions in protein–protein interfaces. *Proteins: Struct., Funct., Bioinf.* **2005,** *59* (2), 231-239.

24. Wang, X.; Sarycheva, O. V.; Koivisto, B. D.; McKie, A. H.; Hof, F., A Terphenyl Scaffold for π-Stacked Guanidinium Recognition Elements. *Org. Lett.* **2008,** *10* (2), 297-300.

25. Thompson, S. E.; Smithrud, D. B., Carboxylates Stacked over Aromatic Rings Promote Salt Bridge Formation in Water. *J. Am. Chem. Soc.* **2002,** *124* (3), 442-449.

26. Wang, X.; Post, J.; Hore, D. K.; Hof, F., Minimalist Synthetic Host with Stacked Guanidinium Ions Mimics the Weakened Hydration Shells of Protein–protein Interaction Interfaces. *The Journal of Organic Chemistry* **2014,** *79* (1), 34-40.

27. Lee, W.-G.; Gallardo-Macias, R.; Frey, K. M.; Spasov, K. A.; Bollini, M.; Anderson, K. S.; Jorgensen, W. L., Picomolar Inhibitors of HIV Reverse Transcriptase Featuring Bicyclic Replacement of a Cyanovinylphenyl Group. *J. Am. Chem. Soc.* **2013,** *135* (44), 16705-16713.

28. Nieto-Draghi, C.; Fayet, G.; Creton, B.; Rozanska, X.; Rotureau, P.; de Hemptinne, J.-C.; Ungerer, P.; Rousseau, B.; Adamo, C., A General Guidebook for the Theoretical Prediction of Physicochemical Properties of Chemicals for Regulatory Purposes. *Chem. Rev. (Washington, DC, U. S.)* **2015,** *115* (24), 13093-13164.

29. Giroud, M.; Ivkovic, J.; Martignoni, M.; Fleuti, M.; Trapp, N.; Haap, W.; Kuglstatter, A.; Benz, J.; Kuhn, B.; Schirmeister, T.; Diederich, F., Inhibition of the Cysteine Protease Human Cathepsin L by Triazine Nitriles: Amide···Heteroarene π-Stacking Interactions and Chalcogen Bonding in the S3 Pocket. *ChemMedChem* **2017,** *12* (3), 257-270.

30. Huber, R. G.; Margreiter, M. A.; Fuchs, J. E.; von Grafenstein, S.; Tautermann, C. S.; Liedl, K. R.; Fox, T., Heteroaromatic π-Stacking Energy Landscapes. *J. Chem. Inf. Model.* **2014,** *54* (5), 1371-1379.

31. An, Y.; Doney, A. C.; Andrade, R. B.; Wheeler, S. E., Stacking Interactions between 9-Methyladenine and Heterocycles Commonly Found in Pharmaceuticals. *J. Chem. Inf. Model.* **2016,** *56* (5), 906-914.

32. Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A., The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Acc. Chem. Res.* **2016,** *49* (6), 1292-1301.

33. Walvoord, R. R.; Huynh, P. N. H.; Kozlowski, M. C., Quantification of Electrophilic Activation by Hydrogen-Bonding Organocatalysts. *J. Am. Chem. Soc.* **2014,** *136* (45), 16055-16065.

34. Orlandi, M.; Coelho, J. A. S.; Hilton, M. J.; Toste, F. D.; Sigman, M. S., Parametrization of Non-covalent Interactions for Transition State Interrogation Applied to Asymmetric Catalysis. *J. Am. Chem. Soc.* **2017,** *139* (20), 6803-6806.

35. Santiago, C. B.; Guo, J.-Y.; Sigman, M. S., Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* **2018,** *9* (9), 2398-2412.

36. Metsanen, T. T.; Lexa, K. W.; Santiago, C. B.; Chung, C. K.; Xu, Y.; Liu, Z.; Humphrey, G. R.; Ruck, R. T.; Sherer, E. C.; Sigman, M. S., Combining traditional 2D and modern physical organic-derived descriptors to predict enhanced enantioselectivity for the key aza-Michael conjugate addition in the synthesis of Prevymis (letermovir). *Chem. Sci.* **2018,** *9* (34), 6922-6927.

37. Santiago, C. B.; Milo, A.; Sigman, M. S., Developing a Modern Approach To Account for Steric Effects in Hammett-Type Correlations. *J. Am. Chem. Soc.* **2016**.

38. Niemeyer, Z. L.; Milo, A.; Hickey, D. P.; Sigman, M. S., Parameterization of phosphine ligands reveals mechanistic pathways and predicts reaction outcomes. *Nat Chem* **2016,** *8* (6), 610-7.

39. Milo, A.; Neel, A. J.; Toste, F. D.; Sigman, M. S., Organic chemistry. A data-intensive approach to mechanistic elucidation applied to chiral anion catalysis. *Science* **2015,** *347* (6223), 737-43.

40. Harper, K. C.; Bess, E. N.; Sigman, M. S., Multidimensional steric parameters in the analysis of asymmetric catalytic reactions. *Nat Chem* **2012,** *4* (5), 366-74.

41. Harper, K. C.; Vilardi, S. C.; Sigman, M. S., Prediction of catalyst and substrate performance in the enantioselective propargylation of aliphatic ketones by a multidimensional model of steric effects. *J. Am. Chem. Soc.* **2013,** *135* (7), 2482-5.

42. Urbano-Cuadrado, M.; Carbo, J. J.; Maldonado, A. G.; Bo, C., New quantum mechanics-based three-dimensional molecular descriptors for use in QSSR approaches: application to asymmetric catalysis. *J. Chem. Inf. Model.* **2007,** *47* (6), 2228-34.

43. Garcia-Lopez, D.; Cid, J.; Marques, R.; Fernandez, E.; Carbo, J. J., Quantitative Structure-Activity Relationships for the Nucleophilicity of Trivalent Boron Compounds. *Chemistry* **2017,** *23* (21), 5066-5075.

44. Piou, T.; Romanov-Michailidis, F.; Romanova-Michaelides, M.; Jackson, K. E.; Semakul, N.; Taggart, T. D.; Newell, B. S.; Rithner, C. D.; Paton, R. S.; Rovis, T., Correlating Reactivity and Selectivity to Cyclopentadienyl Ligand Properties in Rh(III)-Catalyzed C-H Activation Reactions: An Experimental and Computational Study. *J. Am. Chem. Soc.* **2017,** *139* (3), 1296-1310.

45. Houk, K. N.; Cheong, P. H.-Y., Computational prediction of small-molecule catalysts. *Nature* **2008,** *455* (7211), 309.

46. Houk, K. N.; Liu, F., Holy Grails for Computational Organic Chemistry and Biochemistry. *Acc. Chem. Res.* **2017,** *50*, 539-543.

47. Hopmann, K. H., Quantum Chemical Studies of Asymmetric Reactions: Historical Aspects and Recent Examples. *Int. J. Quant. Chem.* **2015,** *115*, 1232-1249.

48. Lam, Y.; Grayson, M. N.; Holland, M. C.; Simon, A.; Houk, K. N., Theory and Modeling of Asymmetric Catalytic Reactions. *Acc. Chem. Res.* **2016,** *49*, 750-762.

49. Maji, R.; Mallojjala, S. C.; Wheeler, S. E., Chiral phosphoric acid catalysis: from numbers to insights. *Chem. Soc. Rev.* **2018**.

50. Jindal, G.; Sunoj, R. B., Rational design of catalysts for asymmetric diamination reaction using transition state modeling *Org. Biomol. Chem.* **2014,** *12*, 2745-2753.

51. Seguin, T. J.; Wheeler, S. E., Stacking and Electrostatic Interactions Drive the Stereoselectivity of Silylium-Ion Asymmetric Counteranion-Directed Catalysis. *Angew. Chem. Int. Ed.* **2016,** *55* (51), 15889-15893.

52. Guan, Y.; Wheeler, S. E., Automated Quantum Mechanical Predictions of Enantioselectivity in a Rhodium-Catalyzed Asymmetric Hydrogenation. *Angew. Chem. Int. Ed.* **2017,** *56* (31), 9101-9105.

53. Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E., AARON: An Automated Reaction Optimizer for New Catalysts. *J. Chem. Theory Comput.* **2018**.

54. Le, P. Q.; Nguyen, T. S.; May, J. A., A General Method for the Enantioselective Synthesis of α-Chiral Heterocycles. *Org. Lett.* **2012,** *14* (23), 6104-6107.

55. Hunter, C. A.; Sanders, J. K. M., The Nature of π-π Interactions. *J. Am. Chem. Soc.* **1990,** *112*, 5525-5534.

56. Bloom, J. W. G.; Wheeler, S. E., Taking Aromaticity out of Aromatic Interactions. *Angew. Chem. Int. Ed.* **2011,** *50*, 7847-7849.

57. Bissantz, C.; Kuhn, B.; Stahl, M., A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010,** *53*, 5061-5084.

58. Meyer, E. A.; Castellano, R. K.; Diederich, F., Interactions with Aromatic Rings in Chemical and Biological Recognition. *Angew. Chem. Int. Ed.* **2003,** *42*, 1210-1250.

59. Schneider, H.-J., Binding Mechanisms in Supramolecular Complexes. *Angew. Chem. Int. Ed.* **2009,** *48*, 3924-3977.

60. Salonen, L. M.; Ellermann, M.; Diederich, F., Aromatic Rings in Chemical and Biological Recognition: Energetics and Structures. *Angew. Chem. Int. Ed.* **2011,** *50*, 4808-4842.

61. Persch, E.; Dumele, O.; Diederich, F., Molecular Recognition in Chemical and Biological Systems. *Angew. Chem. Int. Ed.* **2015,** *54*, 3290-3327.

62. Wheeler, S. E.; Seguin, T. J.; Guan, Y.; Doney, A. C., Non-covalent Interactions in Organocatalysis and the Prospect of Computational Catalyst Design. *Acc. Chem. Res.* **2016,** *49*, 1061-1069.

63. Toste, F. D.; Sigman, M. S.; Miller, S. J., Pursuit of Noncovalent Interactions for Strategic Site-Selective Catalysis. *Acc. Chem. Res.* **2017,** *50* (3), 609-615.

64. Neel, A. J.; Hilton, M. J.; Sigman, M. S.; Toste, F. D., Exploiting non-covalent π interactions for catalyst design. *Nature* **2017,** *543*, 637-646.

65. Imai, Y. N.; Inoue, Y.; Nakanishi, I.; Kitaura, K., Amide–π interactions between formamide and benzene. *J. Comput. Chem.* **2009,** *30* (14), 2267-2276.

66. Bendová, L.; Jurečka, P.; Hobza, P.; Vondrášek, J., Model of Peptide Bond−Aromatic Ring Interaction: Correlated Ab Initio Quantum Chemical Study. *J. Phys. Chem. B* **2007,** *111* (33), 9975-9979.

67. Hunter, C. A.; Lawson, K. R.; Perkins, J.; Urch, C. J., Aromatic Interactions. *J. Chem. Soc., Perkin Trans. 2* **2001**, 651-669.

68. Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D., Estimates of the Ab Initio Limit for π−π Interactions: The Benzene Dimer. *J. Am. Chem. Soc.* **2002,** *124*, 10887-10893.

69. Wheeler, S. E.; Houk, K. N., Substituent Effects in the Benzene Dimer are Due to Direct Interactions of the Substituents with the Unsubstituted Benzene. *J. Am. Chem. Soc.* **2008,** *130* (33), 10854-10855.

70. Grimme, S., Do Special Noncovalent π-π Stacking Interactions Really Exist? *Angew. Chem. Int. Ed.* **2008,** *47*, 3430-3434.

71. Raju, R. K.; Bloom, J. W. G.; An, Y.; Wheeler, S. E., Substituent Effects on Non-Covalent Interactions with Aromatic Rings: Insights from Computational Chemistry. *ChemPhysChem* **2011,** *12* (17), 3116-3130.

72. Wheeler, S. E., Local Nature of Substituent Effects in Stacking Interactions. *J. Am. Chem. Soc.* **2011,** *133* (26), 10262-10274.

73. Martinez, C. R.; Iverson, B. L., Rethinking the Term "π-Stacking". *Chem. Sci.* **2012,** *3*, 2191-2201.

74. Raju, R. K.; Bloom, J. W. G.; Wheeler, S. E., Broad Transferability of Substituent Effects in π-Stacking Interactions Provides New Insights into Their Origin. *J. Chem. Theory Comput.* **2013,** *9* (8), 3479-3490.

75. Wheeler, S. E., Understanding Substituent Effects in Noncovalent Interactions Involving Aromatic Rings. *Acc. Chem. Res.* **2013,** *46* (4), 1029-1038.

76. Wheeler, S. E.; Bloom, J. W. G., Toward a More Complete Understanding of Noncovalent Interactions Involving Aromatic Rings. *J. Phys. Chem. A* **2014,** *118* (32), 6133-6147.

77. Vitaku, E.; Smith, D. T.; Njardarson, J. T., Analysis of the Structural Diversity, Substitution Patterns, and Frequency of Nitrogen Heterocycles among U.S. FDA Approved Pharmaceuticals. *J. Med. Chem.* **2014,** *57* (24), 10257-10274.

78. Ertl, P.; Jelfs, S.; Mühlbacher, J.; Schuffenhauer, A.; Selzer, P., Quest for the Rings. In Silico Exploration of Ring Universe To Identify Novel Bioactive Heteroaromatic Scaffolds. *J. Med. Chem.* **2006,** *49* (15), 4568-4573.

79. Broughton, H. B.; Watson, I. A., Selection of heterocycles for drug design. *J. Mol. Graph. Model.* **2004,** *23* (1), 51-58.

80. Shi, Y.; Sitkoff, D.; Zhang, J.; Klei, H. E.; Kish, K.; Liu, E. C. K.; Hartl, K. S.; Seiler, S. M.; Chang, M.; Huang, C.; Youssef, S.; Steinbacher, T. E.; Schumacher, W. A.; Grazier, N.; Pudzianowski, A.; Apedo, A.; Discenza, L.; Yanchunas, J.; Stein, P. D.; Atwal, K. S., Design, Structure−Activity Relationships, X-ray Crystal Structure, and Energetic Contributions of a Critical P1 Pharmacophore: 3-Chloroindole-7-yl-Based Factor Xa Inhibitors. *J. Med. Chem.* **2008,** *51* (23), 7541-7551.

81. We note that the introduction of nitrogens into the ring of a heterocycle does not actually deplete the pi-electron density. See Wheeler and Bloom, Chem. Commun. 50, 11118 (2014).,

82. Giroud, M.; Harder, M.; Kuhn, B.; Haap, W.; Trapp, N.; Schweizer, W. B.; Schirmeister, T.; Diederich, F., Fluorine Scan of Inhibitors of the Cysteine Protease Human Cathepsin L: Dipolar and Quadrupolar Effects in the π-Stacking of Fluorinated Phenyl Rings on Peptide Amide Bonds. *ChemMedChem* **2016,** *11* (10), 1042-1047.

83. Ehmke, V.; Winkler, E.; Banner, D. W.; Haap, W.; Schweizer, W. B.; Rottmann, M.; Kaiser, M.; Freymond, C.; Schirmeister, T.; Diederich, F., Optimization of Triazine Nitriles as Rhodesain Inhibitors: Structure–Activity Relationships, Bioisosteric Imidazopyridine Nitriles, and X-ray Crystal Structure Analysis with Human Cathepsin L. *ChemMedChem* **2013,** *8* (6), 967-975.

84. Diederich et al. classified local stacked minima into sandwich and parallel-displaced dimer configurations based on the relative lateral displacement of the heterocycle centroid relative to the amide carbonyl.

85. Chai, J.-D.; Head-Gordon, M., Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **2008,** *128* (8), 084106.

86. Burns, L. A.; Vázquez-Mayagoitia, Á.; Sumpter, B. G.; Sherrill, C. D., Density-functional approaches to noncovalent interactions: A comparison of dispersion corrections (DFT-D), exchange-hole dipole moment (XDM) theory, and specialized functionals. *J. Chem. Phys.* **2011,** *134* (8), 084107.

87. Pinski, P.; Riplinger, C.; Valeev, E. F.; Neese, F., Sparse maps—A systematic infrastructure for reduced-scaling electronic structure methods. I. An efficient and simple linear scaling local MP2 method that uses an intermediate basis of pair natural orbitals. *J. Chem. Phys.* **2015,** *143* (3), 034108.

88. Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F., Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *J. Chem. Phys.* **2013,** *139* (13), 134101.

89. Riplinger, C.; Neese, F., An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.* **2013,** *138* (3), 034106.

90. Neese, F.; Hansen, A.; Liakos, D. G., Efficient and accurate approximations to the local coupled cluster singles doubles method using a truncated pair natural orbital basis. *J. Chem. Phys.* **2009,** *131* (6), 064103.

91. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É., Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research*, 2011; Vol. 12, pp 2825-2830.

92. An, Y.; Bloom, J. W. G.; Wheeler, S. E., Quantifying the π-Stacking Interactions in Nitroarene Binding Sites of Proteins. *J. Phys. Chem. B* **2015,** *119* (45), 14441-14450.

93. Pennington, L. D.; Moustakas, D. T., The Necessary Nitrogen Atom: A Versatile High-Impact Design Element for Multiparameter Optimization. *J. Med. Chem.* **2017,** *60* (9), 3552-3579.

94. Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H.; Izmaylov, A.; Bloino, J.; Zheng, G.; Sonnenberg, J.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J., JA; Peralta, J.; Ogliaro, F.; Bearpark, M.; Heyd, J.; Brothers, E.; Kudin, K.; Staroverov, V.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J.; Iyengar, S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J.; Klene, M.; Knox, J.; Cross, J.; Bakken, V.; Adamo, C.; ramillo, J.; Gomperts, R.; Stratmann, R.; Yazyev, O.; Austin, A.; Cammi, R.; Pomelli, C.; Ochterski, J.; Martin, R.; Morokuma, K.; Zakrzewski, V.; Voth, G.; Salvador, P.; Dannenberg, J.; Dapprich, S.; Daniels, A.; Farkas, O.; Foresman, J.; Ortiz, J.; Cioslowski, J.; Fox, D. *Gaussian 09, Revision D.01*, Gaussian, Inc.: 2009.

95. Neese, F., Software update: the ORCA program system, version 4.0. *WIREs Comput. Mol. Sci.* **2017,** *doi: 10.1002/wcms.1327*.

96. Meyer, E. A.; Castellano, R. K.; Diederich, F., Interactions with Aromatic Rings in Chemical and Biological Recognition. *Angew. Chem. Int. Ed.* **2003,** *42*, 1210-1250.

97. Wheeler, S. E.; Seguin, T. J.; Guan, Y.; Doney, A. C., Noncovalent Interactions in Organocatalysis and the Prospect of Computational Catalyst Design. *Acc. Chem. Res.* **2016,** *49* (5), 1061-1069.

98. Mitra, M.; Manna, P.; Seth, S. K.; Das, A.; Meredith, J.; Helliwell, M.; Bauza, A.; Choudhury, S. R.; Frontera, A.; Mukhopadhyay, S., Salt-bridge-π (sb-π) interactions at work: associative interactions of sb-π, π-π and anion-π in Cu(ii)-malonate-2-aminopyridine-hexafluoridophosphate ternary system. *CrystEngComm* **2013,** *15* (4), 686-696.

99. Fujisawa, K.; Humbert-Droz, M.; Letrun, R.; Vauthey, E.; Wesolowski, T. A.; Sakai, N.; Matile, S., Ion Pair−π Interactions. *J. Am. Chem. Soc.* **2015,** *137* (34), 11047-11056.

100. Fujisawa, K.; Beuchat, C.; Humbert-Droz, M.; Wilson, A.; Wesolowski, T. A.; Mareda, J.; Sakai, N.; Matile, S., Anion–π and Cation–π Interactions on the Same Surface. *Angew. Chem. Int. Ed.* **2014,** *53* (42), 11266-11269.

101. Cotelle, Y.; Benz, S.; Avestro, A. J.; Ward, T. R.; Sakai, N.; Matile, S., Anion–π Catalysis of Enolate Chemistry: Rigidified Leonard Turns as a General Motif to Run Reactions on Aromatic Surfaces. *Angew. Chem. Int. Ed.* **2016,** *55* (13), 4275-4279.

102. Alegre-Requena, J. V.; Marqués-López, E.; Herrera, R. P., "Push–Pull π+/π–" (PPππ) Systems in Catalysis. *ACS Catal.* **2017,** *7* (10), 6430-6439.

103. Kuwada, T.; Hasegawa, T.; Takagi, T.; Sato, I.; Shishikura, F., pH-dependent structural changes in haemoglobin component V from the midge larva Propsilocerus akamusi (Orthocladiinae, Diptera). *Acta crystallographica. Section D, Biological crystallography* **2010,** *66* (Pt 3), 258-67.

104. Prates, J. A. M.; Tarbouriech, N.; Charnock, S. J.; Fontes, C. M. G. A.; Ferreira, L. s. M. A.; Davies, G. J., The Structure of the Feruloyl Esterase Module of Xylanase 10B from Clostridium thermocellum Provides Insights into Substrate Recognition. *Structure* **2001,** *9* (12), 1183-1190.

105. Renault, L.; Kuhlmann, J.; Henkel, A.; Wittinghofer, A., Structural Basis for Guanine Nucleotide Exchange on Ran by the Regulator of Chromosome Condensation (RCC1). *Cell* **2001,** *105* (2), 245-255.

106. Bootsma, A. N.; Wheeler, S. E., Stacking Interactions of Heterocyclic Drug Fragments with Protein Amide Backbones. *ChemMedChem* **2018,** *13*, 835-841.

107. The interaction of a point dipole with a molecule is the dot product of the dipole and the electric field arising from the molecule at the position of the dipole. Thus, dipole-field interactions are maximized when the dipole and field are aligned.

108. Vitaku, E.; Smith, D.; Njardarson, J., Analysis of the Structural Diversity, Substitution Patterns, and Frequency of Nitrogen Heterocycles among U.S. FDA Approved Pharmaceuticals. *J. Med. Chem.* **2014,** *57* (24), 10257-10274.

109. Broughton, H.; Watson, I., Selection of heterocycles for drug design. *J. Mol. Graphics Modell.* **2004,** *23* (1), 51-58.

110. Chai, J.-D.; Head-Gordon, M., Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008,** *10* (44), 6615-6620.

111. Weigend, F.; Ahlrichs, R., Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005,** *7*, 3297-3305.

112. Dunning, T. H., Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989,** *90* (2), 1007-1023.

113. Zhang, X.; DeChancie, J.; Gunaydin, H.; Chowdry, A. B.; Clemente, F. R.; Smith; Handel, T. M.; Houk, K. N., Quantum Mechanical Design of Enzyme Active Sites. *The Journal of Organic Chemistry* **2008,** *73* (3), 889-899.

114. Tantillo, D. J.; Jiangang, C.; Houk, K. N., Theozymes and compuzymes: theoretical models for biological catalysis. *Curr. Opin. Chem. Biol.* **1998,** *2* (6), 743-750.

115. Kumar, K.; Woo, S. M.; Siu, T.; Cortopassi, W. A.; Duarte, F.; Paton, R. S., Cation–π interactions in protein–ligand binding: theory and data-mining reveal different roles for lysine and arginine. *Chem. Sci.* **2018,** *9* (10), 2655-2665.

116. Jeziorski, B.; Moszyński, R.; Szalewicz, K., Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van der Waals Complexes. *Chem. Rev.* **1994,** *94*, 1887-1930.

117. Szalewicz, K., Symmetry-Adapted Perturbation Theory of Intermolecular Forces. *WIREs Comp. Mol. Sci.* **2012,** *2*, 254-272.

118. Hohenstein, E. G.; Sherrill, C. D., Density Fitting of Intramonomer Correlation Effects in Symmetry-Adapted Perturbation Theory *J. Chem. Phys.* **2010,** *133*, 014101.

119. Hohenstein, E. G.; Sherrill, C. D., Density Fitting and Cholesky Decomposition Approximations in Symmetry-Adapted Perturbation Theory: Implementation and Application to Probe the Nature of π-π Interactions in Linear Acenes. *J. Chem. Phys.* **2010,** *132*, 184111.

120. Turney, J. M.; Simmonett, A. C.; Parrish, R. M.; Hohenstein, E. G.; Evangelista, F. A.; Fermann, J. T.; Mintz, B. J.; Wilke, J. J.; Abrams, M. L.; Russ, N. J.; Leininger, M. L.; Janssen, C. L.; Seidl, E. T.; Allen, W. D.; Schaefer, H. F.; King, R. A.; Valeev, E. F.;

Sherrill, C. D.; Crawford, T. D., Psi4: An Open-Source *ab Initio* Electronic Structure Program. *WIREs Comp. Mol. Sci.* **2012,** *2*, 556-565.

121. Bijina, P. V.; Suresh, C. H.; Gadre, S. R., Electrostatics for probing lone pairs and their interactions. *J. Comput. Chem.* **2018,** *39* (9), 488-499.

122. Suresh, C. H.; Gadre, S. R., Electrostatic Potential Minimum of the Aromatic Ring as a Measure of Substituent Constant. *J. Phys. Chem. A* **2007,** *111* (4), 710-714.

123. Cortopassi, W. A.; Kumar, K.; Paton, R. S., Cation–π interactions in CREBBP bromodomain inhibition: an electrostatic model for small-molecule binding affinity and selectivity. *Org. Biomol. Chem.* **2016,** *14* (46), 10926-10938.

124. Cockroft, S. L.; Perkins, J.; Zonta, C.; Adams, H.; Spey, S. E.; Low, C. M. R.; Vinter, J. G.; Lawson, K. R.; Urch, C. J.; Hunter, C. A., Substituent effects on aromatic stacking interactions. *Org. Biomol. Chem.* **2007,** *5* (7), 1062-1080.

125. This is a different definition than we have used for similar descriptors in the past, where a square grid was employed. (see ref. 27)

126. Mecozzi, S.; West, A. P., Jr.; Dougherty, D. A., Cation-pi Interactions in Simple Aromatics: Electrostatics Provide a Predictive Tool. *J. Am. Chem. Soc.* **1996,** *118*, 2307-2308.

127. Mecozzi, S.; West, A. P., Jr.; Dougherty, D. A., Cation-pi interactions in aromatics of biological and medicinal interest: electrostatic potential surfaces as a useful qualitative guide. *Proc. Natl. Acad. Sci. USA* **1996,** *93*, 10566-10571.

128. Cubero, E.; Luque, F. J.; Orozco, M., Is polarization important in cation-pi interactions? *Proc. Natl. Acad. Sci. USA* **1998,** *95*, 5976-5980.

129. Kim, D.; Tarakeshwar, P.; Kim, K. S., Theoretical Investigations of Anion−π Interactions: The Role of Anions and the Nature of π Systems. *J. Phys. Chem. A* **2004,** *108* (7), 1250-1258.

130. Garau, C.; Frontera, A.; Quiñonero, D.; Ballester, P.; Costa, A.; Deyà, P. M., Cation−π versus Anion−π Interactions: Energetic, Charge Transfer, and Aromatic Aspects. *J. Phys. Chem. A* **2004,** *108* (43), 9423-9427.

131. Sinnokrot, M. O.; Sherrill, C. D., High-Accuracy Quantum Mechanical Studies of π−π Interactions in Benzene Dimers. *J. Phys. Chem. A* **2006,** *110* (37), 10656-10668.

132. The simple model depicted in Figure 1A suggests the potential for cooperativity between the cation-π and anion-π interactions through the polarization of the heterocycle. However, the observed geometries are incompatible with such an effect; moreover, we observe anti-cooperativity between these two interactions, with the total interaction energy of the salt-bridge with the heterocycle being less than the sum of the interaction energies for the separate cation-π and anion-π interactions (see SI Table S4 for pairwise

interaction energies). This is likely due to partial charge transfer across the salt-bridge, which diminishes the favorability of the cation-π interaction.

133. The SAPT0 interaction energies are strongly correlated with the DLPNO-CCSD(T)/def2-TZVP data (R2 = 0.98), validating the use of SAPT0 for such analyses.

134. The correlation of the electrostatic component of the interaction energies with the heterocycle dipole moment is much weaker (R2 = 0.65).

135. Sanders, J. M., Optimal π-Stacking Interaction Energies in Parallel-Displaced Aryl/Aryl Dimers are Predicted by the Dimer Heavy Atom Count. *J. Phys. Chem. A* **2010,** *114* (34), 9205-9211.

136. The ESP and electric field based descriptors in the gas phase and in solution are very similar. See SI Table S2.

137. Martinez, C. R.; Iverson, B. L., Rethinking the Term "π-Stacking". *Chem. Sci.* **2012,** *3*, 2191-2201.

138. Bloom, J. W. G.; Wheeler, S. E., Taking the Aromaticity out of Aromatic Interactions. *Angew. Chem. Int. Ed.* **2011,** *50* (34), 7847-7849.

139. Neel, A. J.; Hilton, M. J.; Sigman, M. S.; Toste, F. D., Exploiting non-covalent pi interactions for catalyst design. *Nature* **2017,** *543* (7647), 637-646.

140. Thakuria, R.; Nath, N. K.; Saha, B. K., The Nature and Applications of π–π Interactions: A Perspective. *Cryst. Growth Des.* **2019**.

141. Sutton, C.; Risko, C.; Brédas, J.-L., Noncovalent Intermolecular Interactions in Organic Electronic Materials: Implications for the Molecular Packing vs Electronic Properties of Acenes. *Chem. Mater.* **2015,** *28* (1), 3-16.

142. Burley, S.; Petsko, G., Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science* **1985,** *229* (4708), 23-28.

143. Chelli, R.; Gervasio, F. L.; Procacci, P.; Schettino, V., Stacking and T-shape Competition in Aromatic−Aromatic Amino Acid Interactions. *J. Am. Chem. Soc.* **2002,** *124* (21), 6133-6143.

144. McGaughey, G. B.; Gagné, M.; Rappé, A. K., π-Stacking Interactions. Alive and Well in Proteins. *J. Biol. Chem.* **1998,** *273* (25), 15458-15463.

145. Clark, M. P.; Ledeboer, M. W.; Davies, I.; Byrn, R. A.; Jones, S. M.; Perola, E.; Tsai, A.; Jacobs, M.; Nti-Addae, K.; Bandarage, U. K.; Boyd, M. J.; Bethiel, R. S.; Court, J. J.; Deng, H.; Duffy, J. P.; Dorsch, W. A.; Farmer, L. J.; Gao, H.; Gu, W.; Jackson, K.; Jacobs, D. H.; Kennedy, J. M.; Ledford, B.; Liang, J.; Maltais, F.; Murcko, M.; Wang, T.; Wannamaker, M. W.; Bennett, H. B.; Leeman, J. R.; McNeil, C.; Taylor, W. P.; Memmott, C.; Jiang, M.; Rijnbrand, R.; Bral, C.; Germann, U.; Nezami, A.; Zhang, Y.;

Salituro, F. G.; Bennani, Y. L.; Charifson, P. S., Discovery of a Novel, First-in-Class, Orally Bioavailable Azaindole Inhibitor (VX-787) of Influenza PB2. *J. Med. Chem.* **2014,** *57* (15), 6668-6678.

146. Severin, C.; Rocha de Moura, T.; Liu, Y.; Li, K.; Zheng, X.; Luo, M., The cap-binding site of influenza virus protein PB2 as a drug target. *Acta Crystallographica Section D* **2016,** *72* (2), 245-253.

147. Ferreira de Freitas, R.; Harding, R. J.; Franzoni, I.; Ravichandran, M.; Mann, M. K.; Ouyang, H.; Lautens, M.; Santhakumar, V.; Arrowsmith, C. H.; Schapira, M., Identification and Structure–Activity Relationship of HDAC6 Zinc-Finger Ubiquitin Binding Domain Inhibitors. *J. Med. Chem.* **2018,** *61* (10), 4517-4527.

148. Wibowo, A. S.; Singh, M.; Reeder, K. M.; Carter, J. J.; Kovach, A. R.; Meng, W.; Ratnam, M.; Zhang, F.; Dann, C. E., Structures of human folate receptors reveal biological trafficking states and diversity in folate and antifolate recognition. *Proceedings of the National Academy of Sciences* **2013,** *110* (38), 15180.

149. Hudson, W. H.; Ortlund, E. A., The structure, function and evolution of proteins that bind DNA and RNA. *Nat. Rev. Mol. Cell Biol.* **2014,** *15*, 749.

150. Kuo, P.-H.; Chiang, C.-H.; Wang, Y.-T.; Doudeva, L. G.; Yuan, H. S., The crystal structure of TDP-43 RRM1-DNA complex reveals the specific recognition for UG- and TG-rich nucleic acids. *Nucleic Acids Res.* **2014,** *42* (7), 4712-4722.

151. Hunter, C. A.; Sanders, J. K. M., The Nature of π-π Interactions. *J. Am. Chem. Soc.* **1990,** *112*, 5525-5534.

152. Sinnokrot, M. O.; Sherrill, C. D., Substituent Effects in π-π Interactions: Sandwich and T-Shaped Configurations. *J. Am. Chem. Soc.* **2004,** *126*, 7690-7697.

153. Cockroft, S. L.; Hunter, C. A.; Lawson, K. R.; Perkins, J.; Urch, C. J., Electrostatic control of aromatic stacking interactions. *J. Am. Chem. Soc.* **2005,** *127* (24), 8594-5.

154. Cockroft, S. L.; Perkins, J.; Zonta, C.; Adams, H.; Spey, S. E.; Low, C. M.; Vinter, J. G.; Lawson, K. R.; Urch, C. J.; Hunter, C. A., Substituent effects on aromatic stacking interactions. *Org. Biomol. Chem.* **2007,** *5* (7), 1062-80.

155. Ringer, A. L.; Sherrill, C. D., Substituent Effects in Sandwich Configurations of Multiply Substituted Benzene Dimers Are Not Solely Governed By Electrostatic Control. *J. Am. Chem. Soc.* **2009,** *131*, 4574-4575.

156. Cockroft, S. L.; Hunter, C. A., Desolvation and substituent effects in edge-to-face aromatic interactions. *Chem Commun (Camb)* **2009,** (26), 3961-3.

157. Watt, M.; Hardebeck, L. K. E.; Kirkpatrick, C. C.; Lewis, M., Face-to-Face Arene-Arene Bidning Energies: Dominated by Dispersion but Predicted by Electrostatic

and Dispersion/Polarizability Substituent Constants. *J. Am. Chem. Soc.* **2011,** *133*, 3854-3862.

158. Muchowska, K. B.; Adam, C.; Mati, I. K.; Cockroft, S. L., Electrostatic modulation of aromatic rings via explicit solvation of substituents. *J. Am. Chem. Soc.* **2013,** *135* (27), 9976-9.

159. Hwang, J.; Li, P.; Carroll, W. R.; Smith, M. D.; Pellechia, P. J.; Shimizu, K. D., Additivity of substituent effects in aromatic stacking interactions. *J. Am. Chem. Soc.* **2014,** *136* (40), 14060-7.

160. Mignon, P.; Loverix, S.; De Proft, F.; Geerlings, P., Influence of Stacking on Hydrogen Bonding: Quantum Chemical Study on Pyridine−Benzene Model Complexes. *J. Phys. Chem. A* **2004,** *108* (28), 6038-6044.

161. Piacenza, M.; Grimme, S., Van der Waals interactions in aromatic systems: structure and energetics of dimers and trimers of pyridine. *ChemPhysChem* **2005,** *6* (8), 1554-8.

162. Swart, M.; van der Wijst, T.; Fonseca Guerra, C.; Bickelhaupt, F. M., Pi-pi stacking tackled with density functional theory. *J. Mol. Model.* **2007,** *13* (12), 1245-57.

163. Itahara, T.; Imaizumi, K., Role of nitrogen atom in aromatic stacking. *J. Phys. Chem. B* **2007,** *111* (8), 2025-32.

164. Hohenstein, E. G.; Sherrill, C. D., Effects of Heteroatoms on Aromatic π−π Interactions: Benzene−Pyridine and Pyridine Dimer. *J. Phys. Chem. A* **2009,** *113* (5), 878-886.

165. Geng, Y.; Takatani, T.; Hohenstein, E. G.; Sherrill, C. D., Accurately characterizing the pi-pi interaction energies of indole-benzene complexes. *J. Phys. Chem. A* **2010,** *114* (10), 3576-82.

166. Karthikeyan, S.; Nagase, S., Origins of the stability of imidazole-imidazole, benzene-imidazole, and benzene-indole dimers: CCSD(T)/CBS and SAPT calculations. *J. Phys. Chem. A* **2012,** *116* (7), 1694-700.

167. Li, P.; Zhao, C.; Smith, M. D.; Shimizu, K. D., Comprehensive experimental study of N-heterocyclic pi-stacking interactions of neutral and cationic pyridines. *J. Org. Chem.* **2013,** *78* (11), 5303-13.

168. Jacobs, M.; Greff Da Silveira, L.; Prampolini, G.; Livotto, P. R.; Cacelli, I., Interaction Energy Landscapes of Aromatic Heterocycles through a Reliable yet Affordable Computational Approach. *J. Chem. Theory Comput.* **2018,** *14* (2), 543-556.

169. Gung, B. W.; Wekesa, F.; Barnes, C. L., Stacking interactions between nitrogen-containing six-membered heterocyclic aromatic rings and substituted benzene: studies in solution and in the solid state. *J. Org. Chem.* **2008,** *73* (5), 1803-8.

170. Bissantz, C.; Kuhn, B.; Stahl, M., A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010,** *53* (14), 5061-5084.

171. Gonthier, J. F.; Steinmann, S. N.; Roch, L.; Ruggi, A.; Luisier, N.; Severin, K.; Corminboeuf, C., pi-Depletion as a criterion to predict pi-stacking ability. *Chem Commun (Camb)* **2012,** *48* (74), 9239-41.

172. Bootsma, A. N.; Wheeler, S. E., Stacking Interactions of Heterocyclic Drug Fragments with Protein Amide Backbones. *ChemMedChem* **2018,** *13* (8), 835-841.

173. Bootsma, A. N.; Wheeler, S. E., Tuning Stacking Interactions between Asp-Arg Salt-Bridges and Heterocyclic Drug Fragments. *J. Chem. Inf. Model.* **2018**.

174. Bloom, J. W. G.; Raju, R. K.; Wheeler, S. E., Physical Nature of Substituent Effects in XH/π Interactions. *J. Chem. Theory Comput.* **2012,** *8* (9), 3167-3174.

175. The above discussion is based on gas-phase stacking interactions. However, we have also computed stacking interaction energies for all unique dimers in diethyl ether (ε = 4.24) as a model of a typical protein environment and refit equation 1 to these data. While the correlation are not quite as strong as seen for the gas-phase data (R2 = 0.81), the overall trends remain unchanged (see SI Figure S4 and S5) and the guidelines presented should apply to stacking interactions in both the gas-phase and dielectric environments.

176. Grimme, S., Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comp. Chem.* **2006,** *27*, 1787-1799.

177. Becke, A., Density-Functional Thermochemistry. V. Systematic Optimization of Exchange-Correlation Functionals. *J. Chem. Phys.* **1997,** *107*, 8554-8560.

178. 1,9,13, and 15 each with both toluene and p-methylphenol

179. Papajak, E.; Truhlar, D. G., Convergent Partially Augmented Basis Sets for Post-Hartree−Fock Calculations of Molecular Properties and Reaction Barrier Heights. *J. Chem. Theory Comput.* **2011,** *7*, 10-18.

180. Poree, C.; Schoenebeck, F., A Holy Grail in Chemistry: Computational Catalyst Design: Feasible or Fiction? *Acc. Chem. Res.* **2017,** *50* (3), 605-608.

181. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J., Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* **2004,** *3* (11), 935-49.

182. Lu, T.; Porterfield, M. A.; Wheeler, S. E., Explaining the Disparate Stereoselectivities of N-Oxide Catalyzed Allylations and Propargylations of Aldehydes. *Org. Lett.* **2012,** *14* (20), 5310-5313.

183. Zou, Y.; Gutierrez, O.; Sader, A. C.; Patel, N. D.; Fandrick, D. R.; Busacca, C. A.; Fandrick, K. R.; Kozlowski, M.; Senanayake, C. H., A Computational Investigation of the Ligand-Controlled Cu-Catalyzed Site-Selective Propargylation and Allenylation of Carbonyl Compounds. *Org. Lett.* **2017**.

184. Nielsen, M. C.; Bonney, K. J.; Schoenebeck, F., Computational ligand design for the reductive elimination of ArCF(3) from a small bite angle Pd(II) complex: remarkable effect of a perfluoroalkyl phosphine. *Angew Chem Int Ed Engl* **2014,** *53* (23), 5903-6.

185. Gammack Yamagata, A. D.; Datta, S.; Jackson, K. E.; Stegbauer, L.; Paton, R. S.; Dixon, D. J., Enantioselective desymmetrization of prochiral cyclohexanones by organocatalytic intramolecular Michael additions to alpha,beta-unsaturated esters. *Angew Chem Int Ed Engl* **2015,** *54* (16), 4899-903.

186. Straker, R. N.; Peng, Q.; Mekareeya, A.; Paton, R. S.; Anderson, E. A., Computational ligand design in enantio- and diastereoselective ynamide [5+2] cycloisomerization. *Nat Commun* **2016,** *7*, 10109.

187. Burrows, L. C.; Jesikiewicz, L. T.; Lu, G.; Geib, S. J.; Liu, P.; Brummond, K. M., Computationally Guided Catalyst Design in the Type I Dynamic Kinetic Asymmetric Pauson-Khand Reaction of Allenyl Acetates. *J. Am. Chem. Soc.* **2017,** *139* (42), 15022-15032.

188. Iwamoto, H.; Imamoto, T.; Ito, H., Computational design of high-performance ligand for enantioselective Markovnikov hydroboration of aliphatic terminal alkenes. *Nat Commun* **2018,** *9* (1), 2290.

189. Thomas, A. A.; Speck, K.; Kevlishvili, I.; Lu, Z.; Liu, P.; Buchwald, S. L., Mechanistically Guided Design of Ligands That Significantly Improve the Efficiency of CuH-Catalyzed Hydroamination Reactions. *J. Am. Chem. Soc.* **2018,** *140* (42), 13976-13984.

190. Korenaga, T.; Sasaki, R.; Takemoto, T.; Yasuda, T.; Watanabe, M., Computationally-Led Ligand Modification using Interplay between Theory and Experiments: Highly Active Chiral Rhodium Catalyst Controlled by Electronic Effects and CH–π Interactions. *Adv. Synth. Catal.* **2018,** *360* (2), 322-333.

191. Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; Wiest, O.; Norrby, P.-O., Rapid virtual screening of enantioselective catalysts using CatVS. *Nat. Catal.* **2018**.

192. Hansen, E.; Rosales, A. R.; Tutkowski, B.; Norrby, P.-O.; Wiest, O., Prediction of Stereochemistry using Q2MM. *Acc. Chem. Res.* **2016,** *49* (5), 996.

193. Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E., AARON: An Automated Reaction Optimizer for New Catalysts. *J. Chem. Theory Comput.* **2018,** *14* (10), 5249-5261.

194. Reid, J. P.; Sigman, M. S., Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. *Nat Rev Chem* **2018,** *2* (10), 290-305.

195. Le, P. Q.; Nguyen, T. S.; May, J. A., A General Method for the Enantioselective Synthesis of α-Chiral Heterocycles. *Org. Lett.* **2012,** *14*, 6104-6107.

196. Lundy, B. J.; Jansone-Popova, S.; May, J. A., Enantioselective conjugate addition of alkenylboronic acids to indole-appended enones. *Org. Lett.* **2011,** *13* (18), 4958-61.

197. Paton, R. S.; Goodman, J. M.; Pellegrinet, S. C., Theoretical Study of the Asymmetric Conjugate Alkenylation of Enones Catalyzed by Binaphthols. *The Journal of Organic Chemistry* **2008,** *73* (13), 5078-5089.

198. Pellegrinet, S. C.; Goodman, J. M., Asymmetric Conjugate Addition of Alkynylboronates to Enones: Rationale for the Intriguing Catalysis Exerted by Binaphthols. *J. Am. Chem. Soc.* **2006,** *128* (10), 3116-3117.

199. Paton, R. S.; Goodman, J. M.; Pellegrinet, S. C., Mechanistic Insights into the Catalytic Asymmetric Allylboration of Ketones: Brønsted or Lewis Acid Activation? *Org. Lett.* **2009,** *11* (1), 37-40.

200. Wu, T. R.; Chong, J. M., Asymmetric conjugate alkenylation of enones catalyzed by chiral diols. *J. Am. Chem. Soc.* **2007,** *129* (16), 4908-9.

201. Lou, S.; Moquist, P. N.; Schaus, S. E., Asymmetric allylboration of ketones catalyzed by chiral diols. *J. Am. Chem. Soc.* **2006,** *128* (39), 12660-1.

202. Nguyen, T. S.; Yang, M. S.; May, J. A., Experimental mechanistic insight into the BINOL-catalyzed enantioselective conjugate addition of boronates to enones. *Tetrahedron Lett.* **2015,** *56* (23), 3337-3341.

203. Grant, D. J.; Dixon, D. A.; Camaioni, D.; Potter, R. G.; Christe, K. O., Lewis Acidities and Hydride, Fluoride, and X− Affinities of the BH3−nXn Compounds for (X = F, Cl, Br, I, NH2, OH, and SH) from Coupled Cluster Theory. *Inorg. Chem.* **2009,** *48* (18), 8811-8821.

204. Zhao, S.; Gensch, T.; Murray, B.; Niemeyer, Z. L.; Sigman, M. S.; Biscoe, M. R., Enantiodivergent Pd-catalyzed C-C bond formation enabled through ligand parameterization. *Science* **2018**.

205. Santiago, C. B.; Guo, J. Y.; Sigman, M. S., Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* **2018,** *9* (9), 2398-2412.

206. Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A., The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Acc. Chem. Res.* **2016,** *49* (6), 1292-301.

207. Milo, A.; Bess, E. N.; Sigman, M. S., Interrogating selectivity in catalysis using molecular vibrations. *Nature* **2014,** *507* (7491), 210-4.

208. Kozlowski, M. C.; Dixon, S. L.; Panda, M.; Lauri, G., Quantum mechanical models correlating structure with selectivity: predicting the enantioselectivity of beta-amino alcohol catalysts in aldehyde alkylation. *J. Am. Chem. Soc.* **2003,** *125* (22), 6614-5.

209. Hageman, J. A.; Westerhuis, J. A.; Frühauf, H.-W.; Rothenberg, G., Design and Assembly of Virtual Homogeneous Catalyst Libraries –Towardsin silico Catalyst Optimisation. *Adv. Synth. Catal.* **2006,** *348* (3), 361-369.

210. Verloop, A., *Drug Design VOl. III*. Academic Press: 1976.

211. Miertus, S.; Scrocco, E.; Tomasi, J., Electrostatic Interaction of a Solute with a Continuum - a Direct Utilization of Abinitio Molecular Potentials for the Prevision of Solvent Effects. *Chem. Phys.* **1981,** *55* (1), 117-129.

212. Grimme, S., Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory. *Chem. Eur. J.* **2012,** *18* (32), 9955-9964.

213. Wheeler, S. E.; Houk, K. N., Integration Grid Errors for Meta-GGA-Predicted Reaction Energies: Origin of Grid Errors for the M06 Suite of Functionals. *J. Chem. Theory Comput.* **2010,** *6* (2), 395-404.

214. Jackson, K. E.; Paton, R. S. *bobbypaton/Sterimol*, v. 1.0; 2018.

215. E. D. Glendening, A. E. R., J. E. Carpenter, F. Weinhold, NBO Version 3.1.

216. Foster, J. P.; Weinhold, F., Natural hybrid orbitals. *J. Am. Chem. Soc.* **1980,** *102* (24), 7211-7218.

217. Doney, A. C.; Bootsma, A. N.; Wheeler, S. E.

218. Tyagarajan, S.; Lowden, C. T.; Peng, Z.; Dykstra, K. D.; Sherer, E. C.; Krska, S. W., Heterocyclic Regioisomer Enumeration (HREMS): A Cheminformatics Design Tool. *J. Chem. Inf. Model.* **2015,** *55* (6), 1130-5.

219. Thanthiriwatte, K. S.; Hohenstein, E. G.; Burns, L. A.; Sherrill, C. D., Assessment of the Performance of DFT and DFT-D Methods for Describing Distance Dependence of Hydrogen-Bonded Interactions. *J. Chem. Theory Comput.* **2011,** *7* (1), 88-96.

220. Burns, L. A.; Vazquez-Mayagoitia, A.; Sumpter, B. G.; Sherrill, C. D., Density-functional approaches to noncovalent interactions: a comparison of dispersion corrections (DFT-D), exchange-hole dipole moment (XDM) theory, and specialized functionals. *J. Chem. Phys.* **2011,** *134* (8), 084107.

221. An, Y.; Doney, A. C.; Andrade, R. B.; Wheeler, S. E., Stacking Interactions between 9-Methyladenine and Heterocycles Commonly Found in Pharmaceuticals. *J. Chem. Inf. Model.* **2016,** *56* (5), 906-14.

222. This fit is from Supporting Information of Ref 11.

223. The models developed should also be applicable to larger heterocycles, but here we only consider up to bicyclic systems.

224. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. *J Cheminform* **2011,** *3*, 33.

225. In Ref. 12, we showed that ESPrange provides slightly more accurate stacking interactions than ESPmax. However, because ESPrange depends on both ESPmax and ESPmin, which vary independently with the number and distribution of heteroatoms, we were unable to develop a sufficiently accurate predictive model of ESPrange.

226. Wheeler, S. E.; Bloom, J. W. G., Anion-π interactions and positive electrostatic potentials of N-heterocycles arise from the positions of the nuclei, not changes in the π-electron distribution. *Chem. Commun. (Cambridge, U. K.)* **2014,** *50* (76), 11118-11121.

227. Pitt, W. R.; Parry, D. M.; Perry, B. G.; Groom, C. R., Heteroaromatic rings of the future. *J. Med. Chem.* **2009,** *52* (9), 2952-63.

228. http://www.frostserver.org/smiles2stacking

APPENDIX A

SUPPLEMENTARY INFORMATION RELATED TO STACKING INTERACTIONS

OF HETEROCYCLIC DRUG FRAGMENTS WITH PROTEIN AMIDE BACKBONES

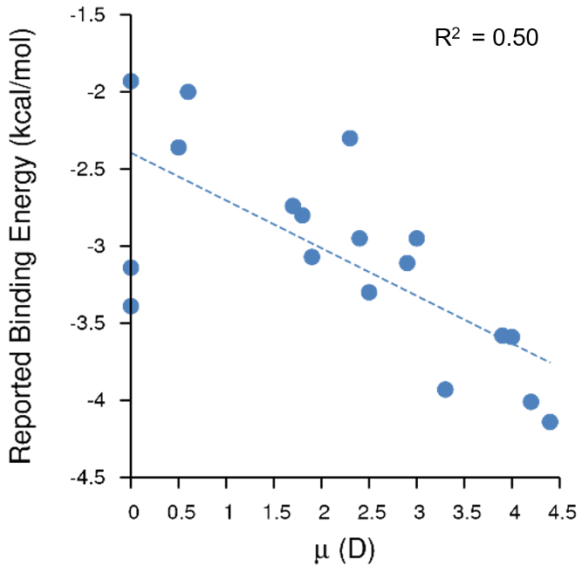


Figure A-1. Correlation between dipole moment and binding energies reported by Diederich et al., adjusted to include heterocycles with zero dipole moment and the global minimum structure for each heterocycle.

Table A-1. Benchmarking of Methods used for three model dimers[a]

	Imidazol-2-one	13	15
HF/aug-ccPVDZ	-2.6	-1.0	0.5
CCSD/aug-cc-pVDZ	-10.4	-7.3	-5.9
CCSD(T)/aug-cc-pVDZ	-12.0	-8.4	-7.1
MP2/aug-cc-pVDZ	-12.7	-8.9	-7.8
HF/aug-cc-pVTZ	-1.7	-0.1	1.5
MP2/aug-cc-pVTZ	-10.8	-7.4	-6.3
SCS-MP2/aug-cc-pVTZ	-8.9	-6.0	-4.9
HF/aug-cc-pVQZ	-1.5	0.1	1.6
MP2/aug-cc-pVQZ	-10.1	-6.7	-5.6
SCS-MP2/aug-cc-pVQZ	-8.1	-5.3	-4.1
HF/aug-cc-pV5Z	-1.4	0.1	1.7
MP2/aug-cc-pV5Z	-9.7	-6.5	-5.4
SCS-MP2/aug-cc-pV5Z	-7.7	-5.0	-3.9
CCSD(T)/CBS [b]	-8.7	-5.8	-4.5
ωb97xd/def2TZVP	-9.1	-6.0	-4.5
ωb97xd/def2QZVP	-8.6	-5.6	-4.2
DLPNO-CCSD(T)/cc-pVTZ	-9.1	-6.2	-4.8
DLPNO-CCSD(T)/cc-pVQZ	-8.7	-5.8	-4.6

[a]all energies computed at the same geometry, optimized at the ωb97xd/def2TZVP level of theory

[b] this value was extrapolated using the focal point method from values above.

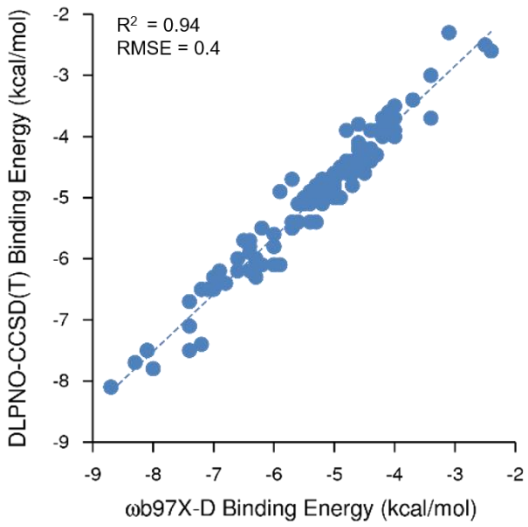


Figure A-2. Correlation between DLPNO-CCSD(T)/cc-pVQZ binding energies and ωb97X-D/def2-TZVP binding energies across all local minima. DLPNO-CCSD(T)/cc-pVQZ energies are single points done at ωb97X-D/def2-TZVP optimized geometries.

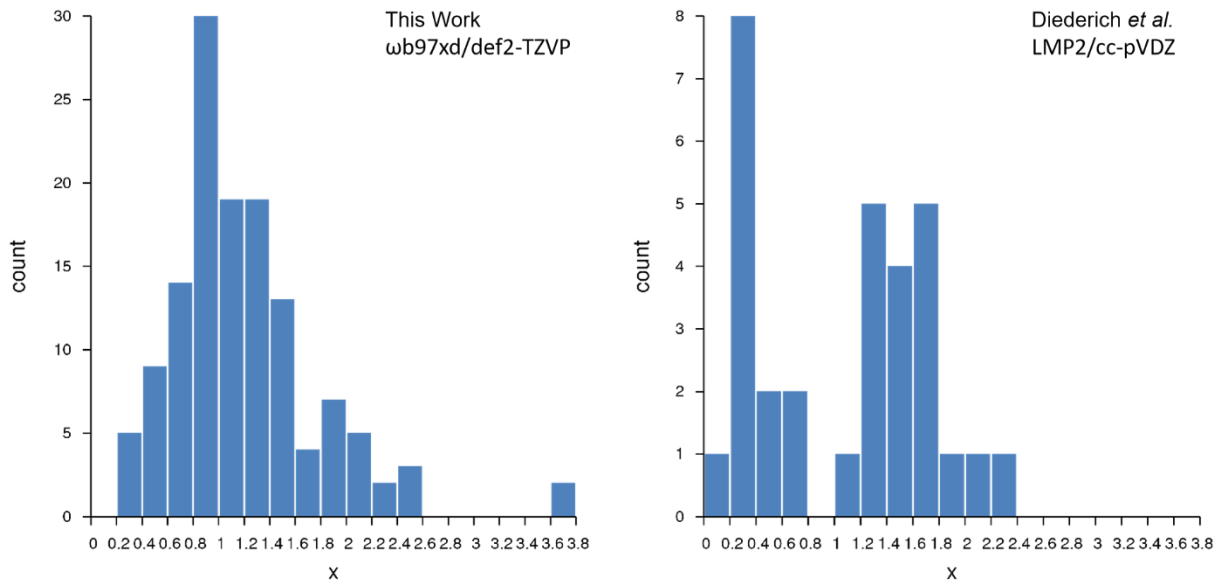


Figure A-3. Histograms showing the distribution of horizontal shifts in this work and in the work presented by Diederich et al., over all reported minima

Table A-2. Comparison of Energies and Geometries for minimum energy structures reported by Diederich et al. and those in this work

		SCS-MP2/aug-ccpVDZ// LMP2/cc-pVDZ [a]				ωB97XD/def2-TZVP			
Heterocycle[b]	**μ (D)**	**E_{bind}**	**d**	**x**	**α**	**E_{bind}**	**d**	**x**	**α**
1	1.9	-3.1	3.70	0.24	166	-6.9	3.17	1.51	172
2	2.3	-2.3	3.68	0.15	125	-6.4	3.24	1.28	124
3	3.9	-3.6	3.50	1.23	173	-7.2	3.22	1.05	154
6	2.9	-3.1	3.36	1.72	158	-7.1	3.24	0.83	180
10	0.6	-2.0	3.71	0.21	158	-5.0	3.23	1.37	66
11	1.7	-2.7	3.45	1.40	131	-5.1	3.25	0.98	174
12	3	-3.0	3.41	1.65	179	-5.4	3.28	0.84	165
13	3.3	-3.9	3.44	0.53	159	-6.0	3.27	0.57	178
14	0.5	-2.4	3.72	0.23	159	-5.0	3.33	1.57	132
15	1.8	-2.8	3.72	0.33	160	-4.8	3.37	1.39	90
18	4	-3.6	3.40	1.47	158	-6.6	3.26	1.14	115
19	0	-1.9	3.81	0.39	--	-5.2	3.34	1.36	--
20	2.4	-3.0	3.70	0.27	159	-5.6	3.33	0.98	152
21	0	-3.1	3.65	0.39	--	-5.7	3.31	0.96	--
22	2.5	-3.3	3.40	1.80	161	-5.9	3.34	0.65	157
23	4.4	-4.1	3.56	0.60	127	-6.6	3.31	0.56	162
24	0	-3.4	3.20	2.09	--	-5.3	3.20	1.50	--
27	4.2	-4.0	3.46	1.78	173	-7.4	3.32	0.81	138

		M06-2X/6-311+G*				B97D/def2-TZVP			
Heterocycle[b]	**μ (D)**	**E_{bind}**	**d**	**x**	**α**	**E_{bind}**	**d**	**x**	**α**
1	1.9	-8.1	3.07	1.37	168	-6.2	3.15	1.53	178
2	2.3	-7.4	3.05	1.33	106	-5.8	3.26	1.18	121
3	3.9	-8.3	3.06	1.07	143	-6.7	3.19	1.09	37
6	2.9	-8.2	3.05	0.92	174	-6.6	3.22	0.86	177
10	0.6	-6.1	3.08	1.32	128	-4.5	3.24	1.42	66
11	1.7	-6.3	3.09	0.95	172	-4.9	3.21	1.40	99
12	3	-6.7	3.10	0.79	160	-5.0	3.29	0.89	147
13	3.3	-7.7	3.09	0.53	171	-5.6	3.31	0.39	169
14	0.5	-6.1	3.20	1.44	133	-4.3	3.34	1.63	127
15	1.8	-6.0	3.24	0.60	175	-4.3	3.47	1.01	86
19	0	-6.2	3.20	1.29	--	-4.6	3.45	1.49	--
20	2.4	-6.5	3.19	0.95	141	-5.1	3.30	1.27	98
21	0	-6.4	3.13	1.02	--	-5.3	3.31	0.99	--
22	2.5	-6.9	3.16	0.89	181	-5.6	3.31	0.87	163
23	4.4	-7.8	3.16	0.65	11	-6.1	3.31	0.59	148
24	0	-6.3	3.00	1.40	--	-5.2	3.19	1.35	--
27	4.2	-8.3	3.12	1.22	78	-7.2	3.16	2.05	153

[a] these values are gathered from parallel displaced and sandwich geometries reported by Diederich et al. taking the minimum energy structure when more than one dimer was reported

[b] only heterocycles that were included in both Diederich et al.'s paper are included in this comparison

APPENDIX B

SUPPLEMENTARY INFORMATION RELATED TO TUNING STACKING INTERACTIONS BETWEEN ASP-ARG SALT BRIDGES AND HETEROCYCLIC DRUG FRAGMENTS

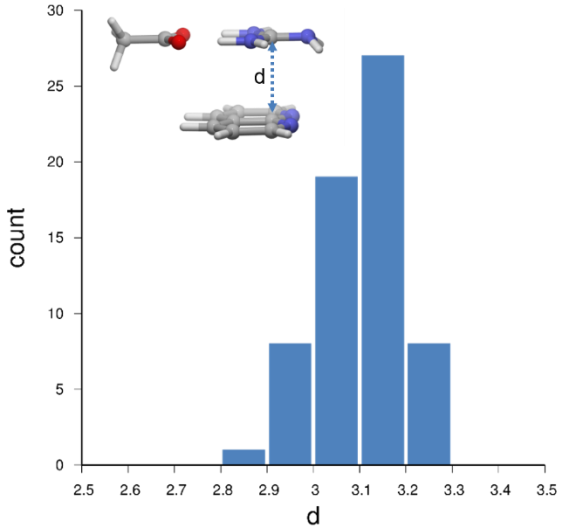


Figure B-1. Distribution of interplanar distances (d, in Angstroms) for all global minimum energy salt-bridge stacked dimers.

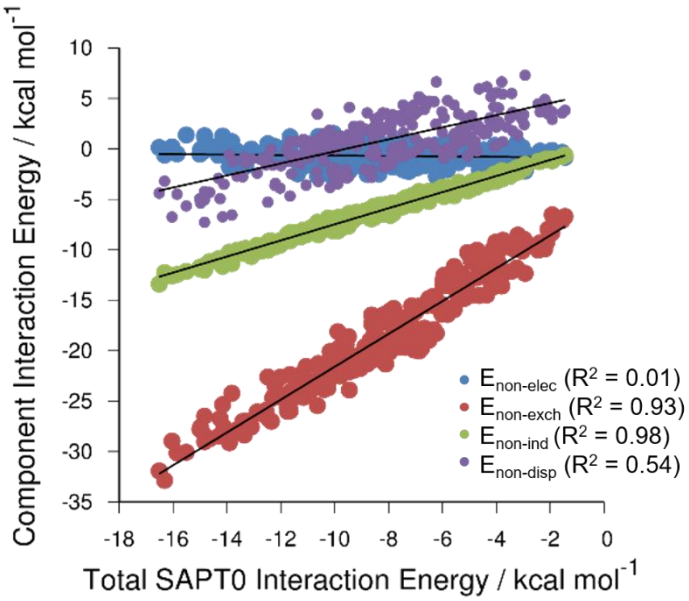
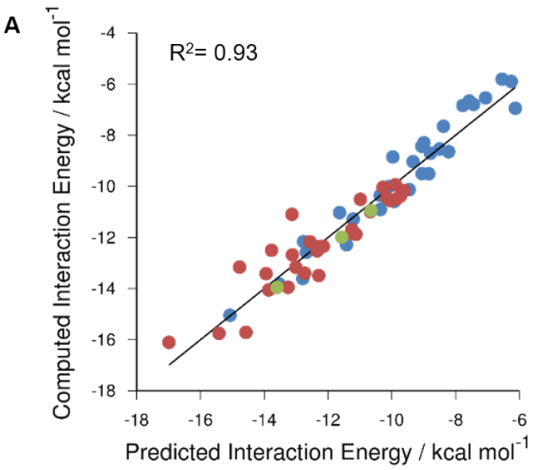
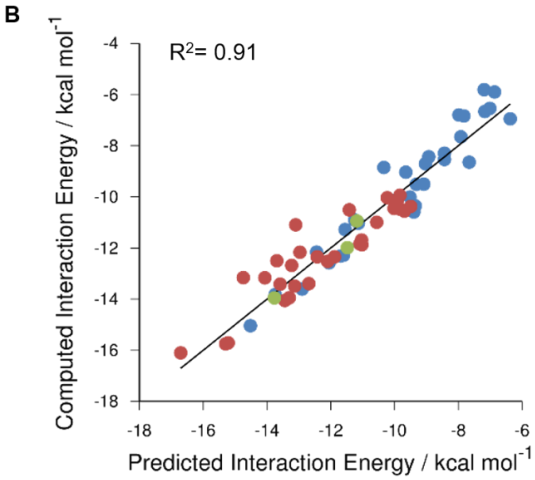


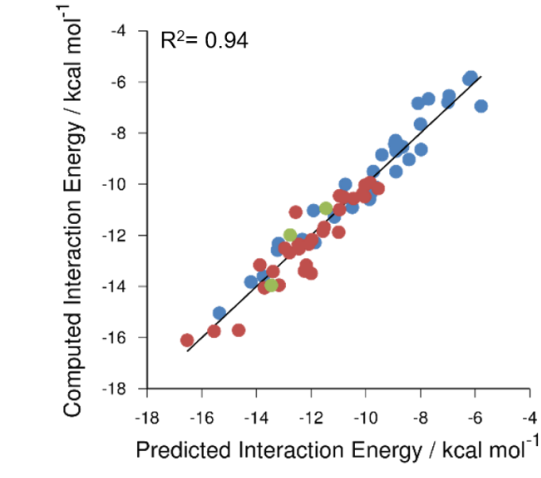
Figure B-2. Non-electrostatic, non-exchange, non-induction, and non-dipersion components versus the total SAPT0 energies for all energy minima.



$$E_{int}^{pred} = -0.46 * ESP_{range}^{gas} + 0.02 * E_{LUMO}^{gas} - 0.08 * \alpha_{zz} - 1.99$$



$$E_{int}^{pred} = -0.41 * ESP_{range}^{gas} - 0.51 * N_{HA} - 1.74$$



$$E_{int}^{pred} = -1.70 * ESP_{\sigma}^{gas} + 0.02 * E_{LUMO}^{gas} - 0.11 * \alpha_{zz} - 1.18$$

Figure B-3. Alternative Predictive Models of Gas Phase Interaction Energies (E_{int})

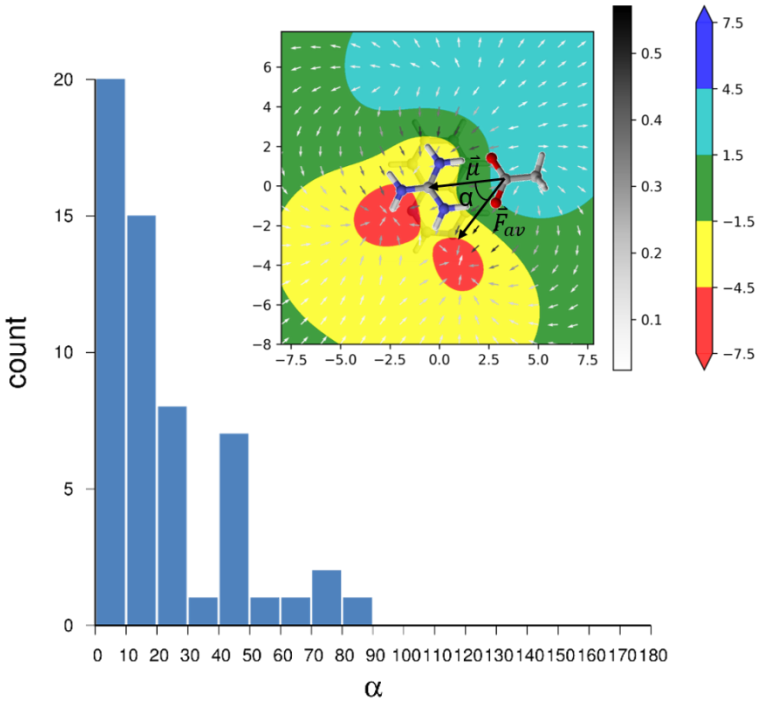


Figure B-4. Distribution of the angle (α, in degrees) between the salt-bridge dipole and the average electric field vector. Note that this angle is not defined for those heterocycles with and average field vector of magntitude 0 (*i.e.* rotationaly symmetric heterocycles)

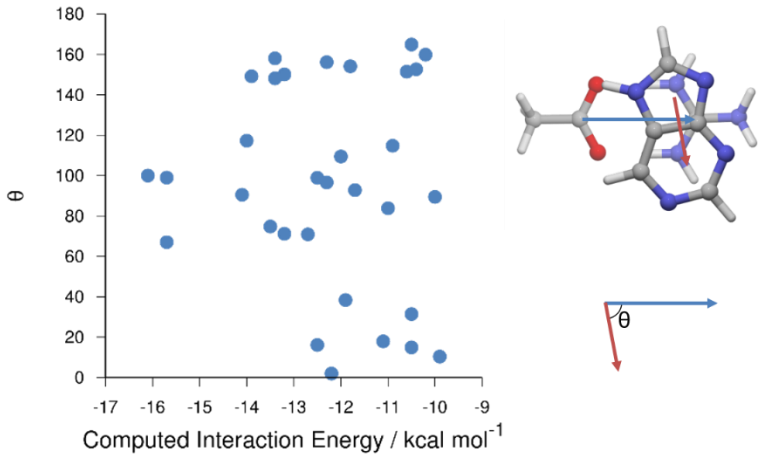


Figure B-5. Orientation of the salt-bridge relative to the heterocycle versus the interaction energy for all global minimum energy structures. Note that this angle is only defined for bicyclic heterocycles

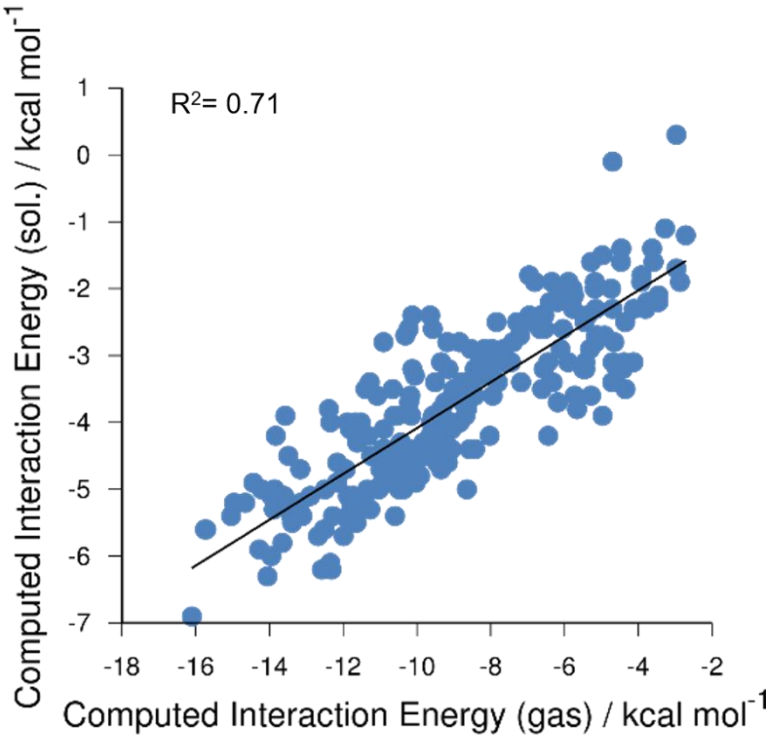
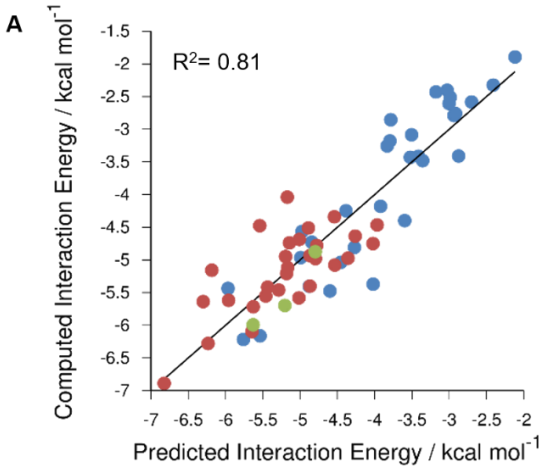
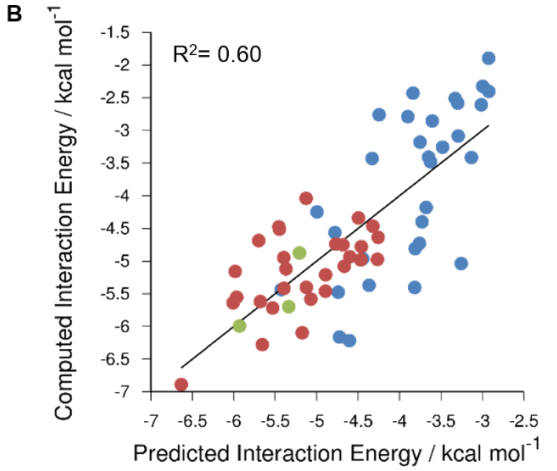


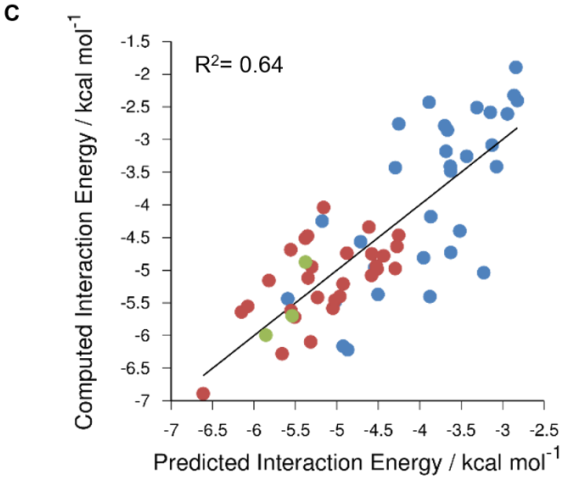
Figure B-6. Comparison of interaction energies computed for all local minima in both the gas phase and the solution phase (ε=4.24)



$$E_{int}^{pred} = -0.11 * ESP_{range}^{sol} + 0.03 * E_{LUMO}^{sol} - 0.04 * \alpha_{zz} - 1.74$$



$$E_{int}^{pred} = -0.09 * ESP_{range}^{sol} - 0.27 * N_{HA} - 0.92$$



$$E_{int}^{pred} = -0.35 * ESP_{\sigma}^{sol} - 0.30 * N_{HA} - 0.69$$

Figure B-7. Alternative Predictive Models of Solution Phase Interaction Energies (E_{int})

APPENDIX C

SUPPLEMENTARY INFORMATION RELATED TO PREDICTING THE STRENGTH OF STACKING INTERACTIONS BETWEEN HETEROCYCLES AND AROMATIC AMINO ACID SIDE CHAINS

SAPT Analysis of all stacked dimers

SAPT enables the decomposition of intermolecular interaction energies into electrostatic (E_{elec}), dispersion (E_{disp}), exchange-repulsion (E_{exch}), and induction (E_{ind}) components, and can be used to identify the effects that drive the trend in interaction energies. Considering all stacked local energy minima for Phe, Tyr, and Trp, the lack of correlation between the total SAPT0 interaction energy (E_{SAPT0}) and both the non-electrostatic component ($E_{non\text{-}elec} = E_{disp} + E_{ind} + E_{exch}$, $R^2 = 0.0$) and the non-dispersion component ($E_{non\text{-}disp} = E_{elec} + E_{ind} + E_{exch}$, $R^2 = 0.02$) indicates that electrostatic and dispersion interactions are vital to capture the overall trends. Moreover, they are of similar importance in determining how strongly a given heterocycle will stack with these aromatic amino acid side chains.

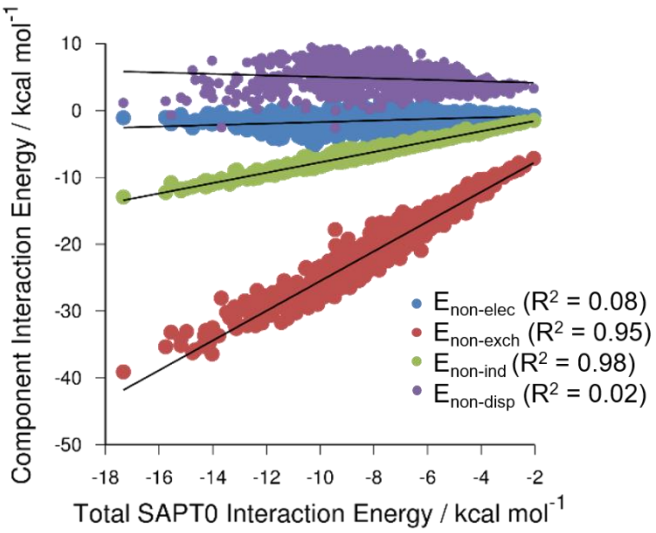


Figure C-1. Non-electrostatic, non-exchange, non-induction, and non-dipersion components versus the total SAPT0 energies for all energy minima.

Other Predictive Models

In addition to equation 4.1 presented in the main text, we also considered other combinations of molecular descriptors that gave rise to fits of similar quality and robustness. Below are plots of two of these fits. The first contains one fewer descriptor that equation 1, and is based only on heavy atom count and the maximum value of the ESP; however it does show slightly decreased performance when compared to equation 4.1.

$$\Delta E_{pred} = N_{HA}^{AA}(-0.03ESP_{max}^{Het} - 0.09N_{HA}^{Het}) - 1.33 \qquad \text{(C-1)}$$

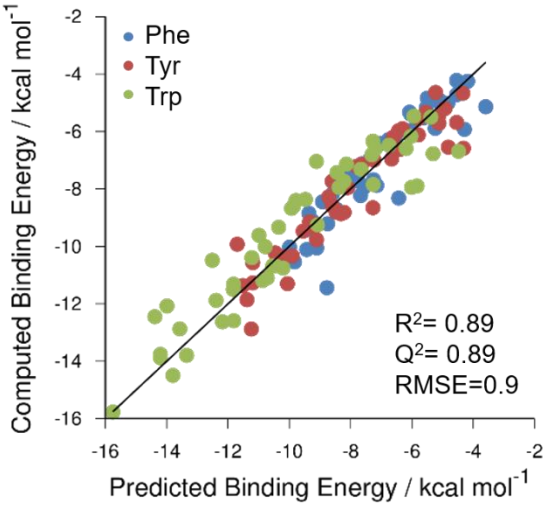


Figure C-2. Scatter plot of computed DLPNO-CCSD(T)/cc-pVQZ interaction energies vs predicted interaction energies from equation (C-1) for global minimum energy stacked dimers of heterocycles **1-47** with Phe, Tyr, and Trp side chains

We also explored fits using the heterocycle dipole moment rather than ESPrange, including

$$\Delta E_{pred} = N_{HA}^{AA}(-0.03\mu - 0.03ESP_{mean}^{Het} - 0.10N_{HA}^{Het}) - 1.66 \quad \text{(C-2)}$$

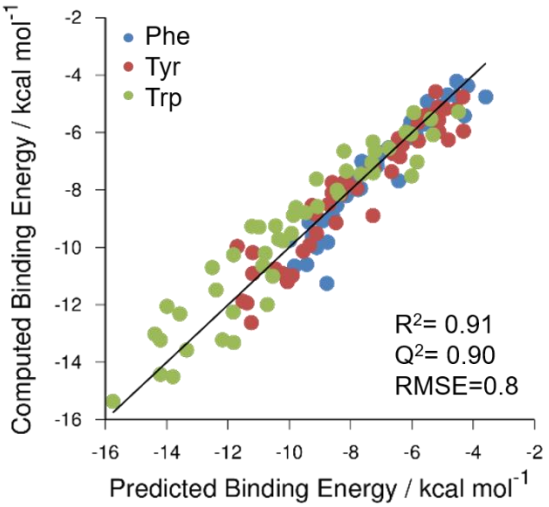


Figure C-3. Scatter plot of computed DLPNO-CCSD(T)/cc-pVQZ interaction energies vs predicted interaction energies from equation (C-2) for global minimum energy stacked dimers of heterocycles **1-47** with Phe, Tyr, and Trp side chains.

Table C-1. Incremental focal point tables for selected stacked dimers (relative energies in kcal mol^{-1}). Extrapolated complete basis set (CBS) limit values are in brackets.

	HF	+δSLMP2	+δDLPNO-CCSD	+δDLPNO-CCSD(T)	= DLPNO-CCSD(T)
			Phe···benzene		
cc-pVDZ	4.18	-10.16	-0.05	-1.1	[-7.13]
cc-pVTZ	6.33	-10.28	0.36	-1.4	[-5.00]
cc-pVQZ	6.7	-10.18	0.49	-1.5	[-4.48]
CBS LIMIT	[+6.76]	[-10.10]	[+0.58]	[-1.56]	[-4.33]
			Phe···pyrazole		
cc-pVDZ	4.38	-9.35	0	-1.05	[-6.02]
cc-pVTZ	6.22	-9.41	0.37	-1.31	[-4.13]
cc-pVQZ	6.64	-9.32	0.52	-1.42	[-3.59]
CBS LIMIT	[+6.75]	[-9.26]	[+0.63]	[-1.50]	[-3.38]
			Phe···3,9-dihydro-purine-2,6-dione		
cc-pVDZ	-3.94	-7.39	0.54	-0.7	[-11.49]
cc-pVTZ	-1.98	-7.77	0.63	-0.79	[-9.91]
cc-pVQZ	-1.37	-7.79	0.59	-0.86	[-9.43]
CBS LIMIT	[-1.12]	[-7.81]	[+0.57]	[-0.91]	[-9.27]
			Tyr···benzene		
cc-pVDZ	3.99	-10.85	0.07	-1.15	[-7.95]
cc-pVTZ	6.37	-11.09	0.49	-1.46	[-5.69]
cc-pVQZ	6.88	-10.91	0.66	-1.55	[-4.92]
CBS LIMIT	[+7.02]	[-10.78]	[+0.77]	[-1.61]	[-4.61]
			Tyr···pyrazole		
cc-pVDZ	4.27	-10.48	0.09	-1.16	[-7.28]
cc-pVTZ	6.45	-10.65	0.46	-1.44	[-5.18]
cc-pVQZ	6.96	-10.54	0.59	-1.56	[-4.54]
CBS LIMIT	[+7.11]	[-10.45]	[+0.69]	[-1.64]	[-4.29]
			Tyr···3,9-dihydro-purine-2,6-dione		
cc-pVDZ	0.78	-13.24	-0.12	-0.92	[-13.50]
cc-pVTZ	3.48	-14.45	-0.21	-1.17	[-12.34]
cc-pVQZ	4.36	-14.52	-0.1	-1.25	[-11.51]
CBS LIMIT	[+4.75]	[-14.56]	[-0.03]	[-1.31]	[-11.16]
			Trp···benzene		
cc-pVDZ	5.37	-13.42	0.18	-1.43	[-9.30]
cc-pVTZ	7.9	-13.48	0.62	-1.73	[-6.70]
cc-pVQZ	8.34	-13.3	0.8	-1.86	[-6.03]
CBS LIMIT	[+8.40]	[-13.17]	[+0.93]	[-1.96]	[-5.80]
			Trp···pyrazole		
cc-pVDZ	4.17	-11.16	-0.14	-1.34	[-8.47]
cc-pVTZ	6.48	-11.14	0.22	-1.56	[-6.00]
cc-pVQZ	6.98	-10.9	0.26	-1.65	[-5.31]
CBS LIMIT	[+7.12]	[-10.73]	[+0.29]	[-1.72]	[-5.04]
			Trp···3,9-dihydro-purine-2,6-dione		
cc-pVDZ	0.6	-16.52	0.81	-1.34	[-16.46]
cc-pVTZ	3.71	-18.11	0.95	-1.67	[-15.12]
cc-pVQZ	4.7	-18.19	1.08	-1.8	[-14.21]
CBS LIMIT	[+5.13]	[-18.24]	[+1.17]	[-1.89]	[-13.83]

APPENDIX D

SUPPLEMENTARY INFORMATION RELATED TO RAPID EVALUATION OF ELECTROSTATIC POTENTIAL BASED HETEROCYCLE DESCRIPTORS

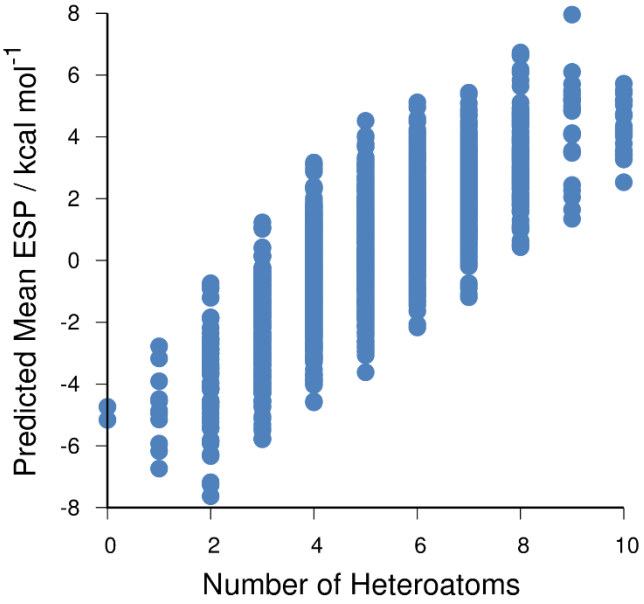


Figure D-1. Trends in ESP_{mean} based on number of heteroatoms in the ring system (C=O is considered a single heteroatom) In general the ESP_{mean} increases with the amount of substitution on the ring, and a very large amount of substitution is required to achieve a high value of ESP_{mean}.

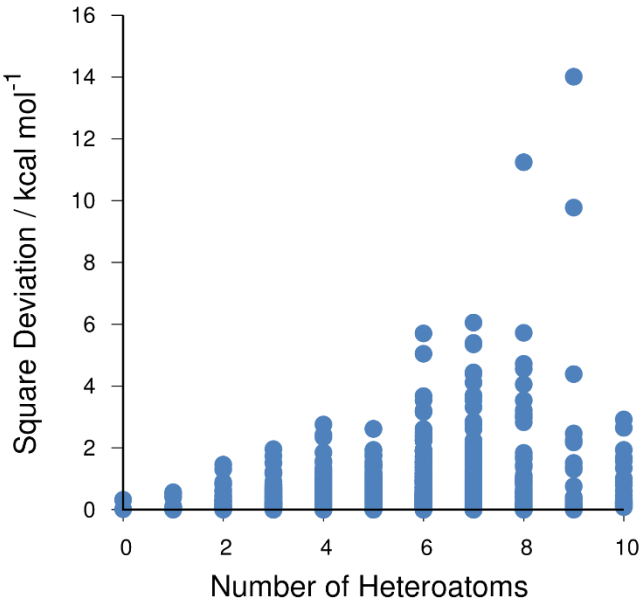
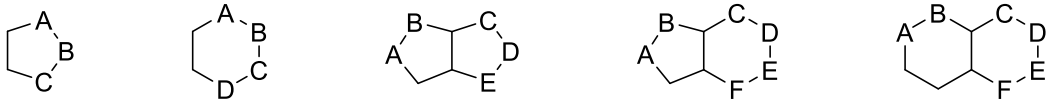


Figure D-2. Trends in error of equation 5.2 based on number of heteroatoms in the ring system. In general increasing amounts of substitution lead to increased errors. This is likely a result of a breakdown in the additivity in substitution effects assumed by equation 5.2 at high levels of substitution. This fact combined with the trend seen in figure D-1 result in in creased errors for large values of ESP_{mean}.

Table D-1. Values for distances and angles used in equation 5.3 Distances are based on a side length of 1, angles in degrees.



5 membered rings			6 membered rings			5-5 bicycles		
pair	**distance**	**angle**	**pair**	**distance**	**angle**	**pair**	**distance**	**angle**
A-B	1	72	**A-B**	1	60	**A-C**	2.60	108
A-C	1.62	144	**A-C**	1.73	120	**A-D**	3.07	180
			A-D	2	180	**B-C**	1.87	36
						B-E	2.50	180

5-6 bicycles			6-6 bicycles		
pair	**distance**	**angle**	**pair**	**distance**	**angle**
A-C	2.60	90	**A-D**	3.48	120
A-D	3.30	150	**A-E**	3.62	180
B-C	1.81	18	**B-C**	1.78	0
B-D	2.69	78	**B-D**	2.68	60
B-E	2.98	138	**B-E**	3.03	120
B-F	2.57	162	**B-F**	2.69	180

APPENDIX E

SUPPLEMENTARY INFORMATION RELATED TO ENANTIOSELECTIVE CATALYST DESIGN THROUGH QUANTUM MECHANICAL VIRTUAL SCREENING

Determining er from ΔΔG‡

There are several different way that enantiomeric ratios can be determined. Most simply it can be calculated from the difference between the lowest R and S transition states (ΔΔG‡)

$$R:S = \frac{e^{\frac{-\Delta\Delta G^{\ddagger}}{RT}}}{1+e^{\frac{-\Delta\Delta G^{\ddagger}}{RT}}} : \frac{1}{1+e^{\frac{-\Delta\Delta G^{\ddagger}}{RT}}} \qquad \text{(E-1)}$$

Consideration of the relative energies for all conformations (i) of both the R ($\Delta G^{\ddagger}_{rel}(R_i)$)and S ($\Delta G^{\ddagger}_{rel}(S_i)$) transition states leads to the following Boltzmann weighted equation for determining *er:*

$$R:S = \frac{\sum_i e^{-\Delta G^{\ddagger}_{rel}(R_i)/RT}}{\sum_i e^{-\Delta G^{\ddagger}_{rel}(R_i)/RT} + \sum_i e^{-\Delta G^{\ddagger}_{rel}(S_i)/RT}} :$$

$$\frac{\sum_i e^{-\Delta G^{\ddagger}_{rel}(S_i)/RT}}{\sum_i e^{-\Delta G^{\ddagger}_{rel}(R_i)/RT} + \sum_i e^{-\Delta G^{\ddagger}_{rel}(S_i)/RT}} \qquad \text{(E-2)}$$

We can further consider the impacts of the relative energy of the racemic background reaction ($\Delta G^{\ddagger}_{rel}(B_i)$)on this Boltzmann weighting :

$$R:S = \frac{\sum_i e^{-\Delta G^{\ddagger}_{rel}(R_i)/RT} + \frac{1}{2}\sum_i e^{-\Delta G^{\ddagger}_{rel}(B_i)/RT}}{\sum_i e^{-\Delta G^{\ddagger}_{rel}(R_i)/RT} + \sum_i e^{-\Delta G^{\ddagger}_{rel}(S_i)/RT} + \sum_i e^{-\Delta G^{\ddagger}_{rel}(B_i)/RT}} :$$

$$\frac{\sum_i e^{-\Delta G^{\ddagger}_{rel}(S_i)/RT} + \frac{1}{2}\sum_i e^{-\Delta G^{\ddagger}_{rel}(B_i)/RT}}{\sum_i e^{-\Delta G^{\ddagger}_{rel}(R_i)/RT} + \sum_i e^{-\Delta G^{\ddagger}_{rel}(S_i)/RT} + \sum_i e^{-\Delta G^{\ddagger}_{rel}(B_i)/RT}} \qquad \text{(E-3)}$$

Table E-1. Predicted *er* values based on inclusion of different conformations and pathways

Catalyst	Exp. *er*	ΔΔG‡	eq. (E-1)	eq. (E-2)	eq. (E-3)
Cat1	73:27	2.12	94:6	92:8	81:19
Cat2	96:4	3.30	99:1	98:2	98:2
Cat3	95:5	3.06	98:2	98:2	97:3
Cat4	97:3	5.17	>99:1	>99:1	>99:1
Cat5	96:4	5.48	>99.1	>99:1	>99:1

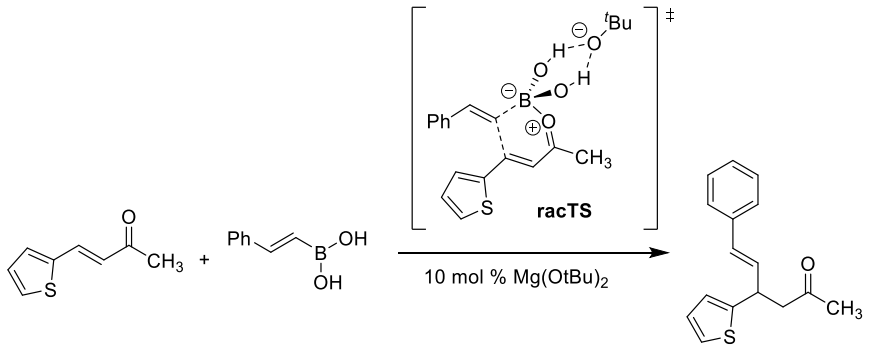


Figure E-1. Racemic background reaction catalyzed by *t*-BuO⁻.

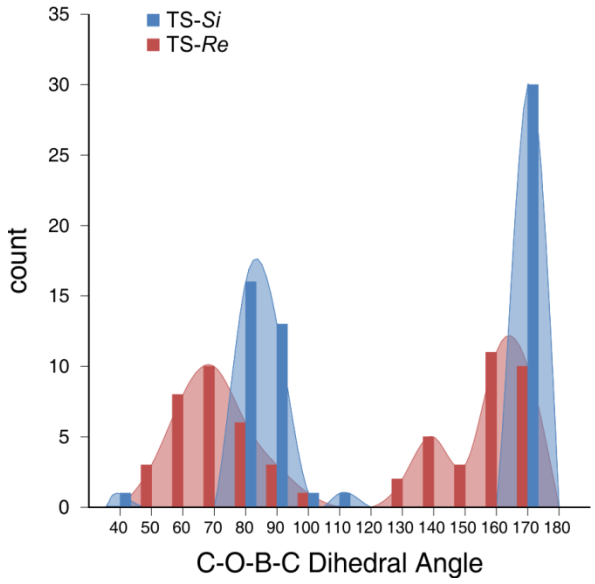


Figure E-2. Distribution of the COBC dihedral angles for TS3-*Re* and TS3-*Si* for all catalysts.