

# **SANDIA REPORT**

SAND2010-6251  
Unlimited Release  
September 2010

## **Multiscale Schemes for the Predictive Description and Virtual Engineering of Materials**

Otto Anatole von Lilienfeld-Toal

Prepared by  
Sandia National Laboratories  
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

**NOTICE:** This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from

U.S. Department of Energy  
Office of Scientific and Technical Information  
P.O. Box 62  
Oak Ridge, TN 37831

Telephone: (865) 576-8401  
Facsimile: (865) 576-5728  
E-Mail: [reports@adonis.osti.gov](mailto:reports@adonis.osti.gov)  
Online ordering: <http://www.osti.gov/bridge>

Available to the public from

U.S. Department of Commerce  
National Technical Information Service  
5285 Port Royal Rd.  
Springfield, VA 22161

Telephone: (800) 553-6847  
Facsimile: (703) 605-6900  
E-Mail: [orders@ntis.fedworld.gov](mailto:orders@ntis.fedworld.gov)  
Online order: <http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online>



SAND2010-6251  
Unlimited Release  
September 2010

# Multiscale Schemes for the Predictive Description and Virtual Engineering of Materials

Otto Anatole von Lilienfeld-Toal  
Multiscale Dynamic Material Modeling Department (1435)  
Sandia National Laboratories  
P.O. Box 5800  
Albuquerque, New Mexico 87185-MS 1322

## Abstract

This report documents research carried out by the author throughout his 3-years Truman fellowship. The overarching goal consisted of developing multiscale schemes which permit not only the predictive description but also the computational design of improved materials. Identifying new materials through changes in atomic composition and configuration requires the use of versatile first principles methods, such as density functional theory (DFT). Using DFT, its predictive reliability has been investigated with respect to pseudopotential construction, band-gap, van-der-Waals forces, and nuclear quantum effects. Continuous variation of chemical composition and derivation of accurate energy gradients in compound space has been developed within a DFT framework for free energies of solvation, reaction energetics, and frontier orbital eigenvalues. Similar variations have been leveraged within classical molecular dynamics in order to address thermal properties of molten salt candidates for heat transfer fluids used in solar thermal power facilities. Finally, a combination of DFT and statistical methods has been used to devise quantitative structure property relationships for the rapid prediction of charge mobilities in polyaromatic hydrocarbons.

## ACKNOWLEDGMENTS

The author greatly acknowledges the outstanding scientific, professional and personal mentorship of Ann E. Mattsson (1435) throughout this project. This report includes work based on collaborations with Sandians, Jean-Loup Faulon, Harry P. Hjalmarson, Sai Jayaraman, Richard B. Lehoucq, Kevin Leung, Milind Misra, Susan B. Rempe, Peter A. Schultz, Aidan P. Thompson, as well as with external scientists, Dennis Andrienko (Max-Planck Institute for Polymer Research, Mainz, Germany), Graeme Henkelman (University of Texas Austin), Alejandro Perez (New York University), Daniel Sheppard (University of Texas Austin), Alexandre Tkatchenko (Fritz-Haber Institute, Berlin, Germany), Mark E. Tuckerman (New York University). The author is grateful for fruitful discussions with Robert W. Bradshaw (1151), Richard B. Diver (1127), Peter J. Feibelman (1415), Nathan P. Siegel (1127). John B. Aidun (Manager of 1435) is acknowledged for endless help and patience.

## CONTENTS

Multiscale Schemes for the Predictive Description and Virtual Engineering of Materials.....	3
Acknowledgments.....	4
Contents.....	5
Nomenclature.....	6
1. Introduction.....	7
2. Accuracy .....	9
3. Alchemical changes and derivatives .....	11
4. Charge mobilities of polyaromatic hydrocarbons.....	13
5. Conclusions.....	14
6. References.....	15
Appendix A: Molten salt eutectics from first principles simulation.....	16
Appendix B: Quantitative structure-property relationships for charge transfer rates of polycyclic aromatic hydrocarbons .....	21
Distribution.....	34

## NOMENCLATURE

AIMD	<i>ab initio</i> molecular dynamics
CCS	Chemical Compound Space
DFT	Density Functional Theory
KS	Kohn-Sham
QSPR	Quantitative Structure Property Relationships

## 1. INTRODUCTION

Enhancing the predictive power of computational materials simulation is instrumental to fulfill the national mission. The virtual tuning of material's properties through the engineering of the atomistic composition and structure represents a potential impact which can hardly be overestimated. Entire technologies, e.g. involving bio-hazards, water purification, explosives, molecular electronics, or harvesting light for renewable energy, would benefit from a successful outcome of such a capability.

Various ingredients are required to achieve said goal.

1. A material's model that is system independent. For this work Density Functional Theory (DFT) was chosen.
2. Sufficiently accurate models yielding material's property predictions which permit the purposeful identification of novel materials with improved properties.
3. A model that allows for variable composition and, if possible, the calculation of gradients in compositional space, dubbed Chemical Compound Space (CCS).
4. Multiscale approaches that allow to address the computational improvement of materials at one time&length scale to optimize properties at other scales.

In the following, progress is documented regarding above points 2 and 3.



## 2. ACCURACY

### 2.1. Band-gap accuracy of DFT

Design of gallium pseudopotentials has been investigated for use in density functional calculations of zinc-blende-type cubic phases of GaAs, GaP, and GaN. A converged construction with respect to all-electron results has been described. Computed lattice constants, bulk moduli, and band gaps vary significantly depending on pseudopotential construction or exchange-correlation functional. The Kohn-Sham band gap of the Ga-(V) semiconductors exhibits a distinctive and strong sensitivity to lattice constant, with near-linear dependence of gap on lattice constant for larger lattice constants and a  $\Gamma$ - $X$  crossover that changes the slope of the dependence. This crossover occurs at  $\approx 98$ , 101, and 95% deviation from the equilibrium lattice constant for GaAs, GaP, and GaN, respectively. See Ref. [1] for details.

### 2.2. Van der Waals forces in DFT

#### 2.2.1. Interatomic many-body contributions in DFT

We have found spuriously large repulsive many-body contributions to binding energies of rare gas systems for the first three rungs of “Jacob’s Ladder” within Kohn-Sham density functional theory. While the description of van der Waals dimers is consistently improved by the pairwise London  $C_6/R^6$  correction, inclusion of a corresponding three-body Axilrod-Teller  $C_9/R^9$  term only increases the repulsive error. Our conclusions, based on extensive solid state and molecular electronic structure calculations, are particularly relevant for condensed phase van der Waals systems. See Ref. [2] for details.

#### 2.2.1. Interatomic 2 and 3-body contributions from DFT

We have presented numerical estimates of the leading two- and three-body dispersion energy terms in van der Waals interactions for a broad variety of molecules and solids. The calculations were based on London and Axilrod–Teller–Muto expressions where the required interatomic dispersion energy coefficients,  $C_6$  and  $C_9$ , are computed “on the fly” from the electron density. Inter- and intramolecular energy contributions have been obtained using the Tang–Toennies (TT) damping function for short interatomic distances. The TT range parameters have equally been extracted on the fly from the electron density using their linear relationship to van der Waals radii. This relationship has empirically been determined for all the combinations of He–Xe rare gas dimers, as well as for the He and Ar trimers. The investigated systems included the S22 database of noncovalent interactions, Ar, benzene and ice crystals, bilayer graphene,  $C_{60}$  dimer, a peptide ( $Ala_{10}$ ), an intercalated drug-DNA model [ellipticine- $d(CG)_2$ ], 42 DNA base pairs, a protein (DHFR, 2616 atoms), double stranded DNA (1905 atoms), and 12 molecular crystal polymorphs from crystal structure prediction blind test studies. The two- and three-body interatomic dispersion energies contribute significantly to binding and cohesive energies, for bilayer graphene the latter reaches 50% of experimentally derived binding energy. These results suggest that interatomic three-body dispersion potentials should be accounted for in atomistic simulations when modeling bulky molecules or condensed phase systems. See Ref. [3] for details.

### 2.3. Nuclear Quantum Effects

Intermolecular enol tautomers of Watson–Crick base pairs could emerge spontaneously via interbase double proton transfer. It has been hypothesized that their formation could be facilitated by thermal fluctuations and proton tunneling, and possibly be relevant to DNA damage. Theoretical and computational studies, assuming classical nuclei, have confirmed the dynamic stability of these rare tautomers. However, by accounting for nuclear quantum effects explicitly through Car–Parrinello path integral molecular dynamics calculations, we have found the tautomeric enol form to be dynamically metastable, with lifetimes too insignificant to be implicated in DNA damage. See Ref. [4] for details.

## 3. ALCHEMICAL CHANGES AND DERIVATIVES

### 3.1. Free Energies of Solvation from Alchemical Growth

We have applied AIMD methods in conjunction with the thermodynamic integration or “ $\lambda$ -path” technique to compute the intrinsic hydration free energies of  $\text{Li}^+$ ,  $\text{Cl}^-$ , and  $\text{Ag}^+$  ions. Using the Perdew–Burke–Ernzerhof functional, adapting methods developed for classical force field applications, and with consistent assumptions about surface potential ( $\phi$ ) contributions, we have obtained absolute AIMD hydration free energies ( $\Delta G_{\text{hyd}}$ ) within a few kcal/mol, or better than 4%, of Tissandier et al.’s [J. Phys. Chem. A 102, 7787 (1998)] experimental values augmented with the SPC/E water model  $\phi$  predictions. The sums of  $\text{Li}^+/\text{Cl}^-$  and  $\text{Ag}^+/\text{Cl}^-$  AIMD  $\Delta G_{\text{hyd}}$ , which are not affected by surface potentials, are within 2.6% and 1.2% of experimental values, respectively. We have also reported the free energy changes associated with the transition metal ion redox reaction  $\text{Ag}^{2+}\text{Ni}^+ \rightarrow \text{Ag}^+\text{Ni}^{2+}$  in water. The predictions for this reaction suggest that existing estimates of  $\Delta G_{\text{hyd}}$  for unstable radiolysis intermediates such as  $\text{Ni}^+$  may need to be extensively revised. See Ref. [5] for details.

### 3.2. Alchemical derivatives of reaction energetics

Based on molecular grand canonical ensemble density functional theory, we have presented a theoretical description of how reaction barriers and enthalpies change as atoms in the system are subjected to alchemical transformations, from one element into another. Changes in the energy barrier for the umbrella inversion of ammonia has been calculated along an alchemical path in which the molecule is transformed into water, and the change in the enthalpy of protonation for methane has been calculated as the molecule is transformed into a neon atom via ammonia, water, and hydrogen fluoride. Alchemical derivatives have been calculated analytically from the electrostatic potential in the unperturbed system, and compared to numerical derivatives calculated with finite difference interpolation of the pseudopotentials for the atoms being transformed. Good agreement has been found between the analytical and numerical derivatives. Alchemical derivatives have thereafter also been shown to be predictive for integer changes in atomic numbers for oxygen binding to a 79 atom palladium nanoparticle, illustrating their potential use in gradient-based optimization algorithms for the rational design of catalysts. See Ref. [6] for details.

### 3.3. Variable Composition and Accurate Derivatives in CCS

Analytical potential energy derivatives, based on the Hellmann–Feynman theorem, have been presented for *any* pair of isoelectronic compounds. Since energies are not necessarily monotonic functions between compounds, these derivatives can fail to predict the right trends of the effect of alchemical mutation. However, quantitative estimates without additional self-consistency calculations can be made when the Hellmann–Feynman derivative is multiplied with a linearization coefficient that is obtained from a reference pair of compounds. These results suggest that accurate predictions can be made regarding any molecule’s energetic properties as long as energies and gradients of three other molecules have been provided. The linearization coefficient can be interpreted as a quantitative measure of chemical similarity. Presented

numerical evidence included predictions of electronic eigenvalues of saturated and aromatic molecular hydrocarbons. See Ref. [7] for details.

### **3.4. Thermal and Transport Properties of Three Alkali Nitrate Salts**

Thermodynamic and transport properties for nitrate salts containing lithium, sodium, and potassium cations were computed from molecular simulations. Densities for the liquid and crystal phases calculated from simulations were within 4% of the experimental values. A nonequilibrium molecular dynamics method was used to compute viscosities and thermal conductivities. The results for the three salts were comparable to the experimental values for both viscosity and thermal conductivity. Computed heat capacities were also in reasonable agreement with experimental values. The computed melting point for  $\text{NaNO}_3$  was within 15 K of its experimental value, while for  $\text{LiNO}_3$  and  $\text{KNO}_3$ , computed melting points were within 100 K of the experimental values. The results show that very small free-energy differences between the crystal and liquid phases can result in large differences in computed melting point. To estimate melting points with an accuracy of around 10 K, simulation methods and force fields must yield free energies with an accuracy of around 0.25 kcal/mol. Tests conducted on a well-studied sodium chloride model indicated negligible dependence of the computed melting point on system size or choice of integration temperature. See Ref. [8] for details.

### **3.5. Molten Salt Eutectics from First Principles Simulation**

Excess free energies of liquid mixtures have been computed using molecular dynamics simulations that involve alchemical transmutations. These free energies, together with free energy differences between liquid and solid pure components as obtained from Ref. [8], have thereafter been combined to determine eutectic compositions and temperatures of the ternary Li, Na, and K nitrate system. See Appendix A for details.

## 4. CHARGE MOBILITIES OF POLYAROMATIC HYDROCARBONS

QSPRs have been developed and assessed for predicting the reorganization energy  $\lambda$  (ranging from 0.1 to 0.3 eV) of polycyclic aromatic hydrocarbons (PAH). Preliminary QSPR models, based on a combination of molecular signature and electronic eigenvalue difference descriptors, have been trained using nearly 200 PAH. Monte Carlo cross-validation systematically improves the performance of the models through progressive reduction of the training set and selection of best performing training subsets. The final biased QSPR model yields correlation coefficients  $q^2$  and  $r^2$  of 0.7 and 0.8, respectively, and an estimated error in predicting  $\lambda$  of +/-0.014 eV. See Appendix B for details.

## 5. CONCLUSIONS

Progress has been documented regarding the quantification of errors made when using pseudopotentials and DFT functionals for band gaps of important semiconductors [1], when using DFT for cohesive energies of weakly bonded systems [2, 3], and when neglecting the nuclear quantum nature in proton transfer reactions [4]. Regarding CCS, studies were carried out dealing with the continuous growth of ions for the computation of free energies of solvation [5], or with the variable composition and derivatives of reaction energetics [6], or with the derivation of general paths and accurate gradients that can connect any two compounds [7]. Using pseudo-critical paths thermal and transport properties of three alkali nitrate salts were studied [8], their ternary eutectics were elucidated through combination of those results with alchemical changes to extract excess free energies of mixing [Appendix A]. Finally, QSPR models were developed for the rapid estimation of reorganization energies of polyaromatic hydrocarbons [Appendix B].

## 6. REFERENCES

1. "Structure and band gaps of Ga-(V) semiconductors: The challenge of Ga pseudopotentials", OAvL and P. A. Schultz, *Phys Rev B* **77** 115202 (2008).
2. "Popular Kohn-Sham density functionals strongly overestimate many-body interactions in van der Waals systems", A. Tkatchenko and OAvL, *Phys. Rev. B* **78** 045116 (2008).
3. "Two and three-body interatomic dispersion energy contributions to binding in molecules and solids", OAvL, A. Tkatchenko, *J. Chem. Phys* **132** 234109 (2010).
4. "Enol tautomers of Watson-Crick base pair models are metastable because of nuclear quantum effects", A. Pérez, M. E. Tuckerman, H. P. Hjalmarson, OAvL, *J. Am. Chem. Soc.* **132** 11510 (2010).
5. "Ab initio molecular dynamics calculations of ion hydration free energies", K. Leung, S. B. Rempe and OAvL, *J. Chem. Phys.* **130** 204507 (2009).
6. "Alchemical derivatives of reaction energetics", D. Sheppard, G. Henkelman, OAvL *J. Chem. Phys.* **133** 084104 (2010).
7. "Accurate ab initio energy gradients in chemical compound space", OAvL, *J. Chem. Phys.* **131** 164102 (2009).
8. "Molecular simulation of the thermal and transport properties of three alkali nitrate salts", S. Jayaraman, A. P. Thompson, OAvL, E. J. Maginn, *Ind. Eng. Chem. Res.* **49** 559 (2010).

# **APPENDIX A: MOLTEN SALT EUTECTICS FROM FIRST PRINCIPLES SIMULATION**

## Molten salt eutectics from first principles simulation

Saivenkataraman Jayaraman,<sup>1</sup> Aidan P. Thompson,<sup>1</sup> and O. Anatole von Lilienfeld<sup>2</sup>

<sup>1</sup>Multiscale Dynamic Materials Modeling Department

<sup>2</sup>Surface and Interface Sciences Department, Sandia National Laboratories, Albuquerque, New Mexico 87185, USA  
(Dated: September 14, 2010)

A method to compute excess free energy of liquid mixtures with molecular dynamics simulations involving alchemical transmutations is presented. These free energies, together with free energy differences between liquid and solid pure components are used to determine eutectic compositions and temperatures.

**Introduction:** Low melting molten salt mixtures are being sought as heat transfer fluids for solar thermal energy systems[1, 2]. Currently, eutectics of alkali and alkaline earth nitrates and nitrites are being used. Due to the high dimensionality of the search space, scanning for low melting mixture compositions experimentally is a challenge. Therefore, eutectics are of great importance in the rational design of optimized materials.

Here, we present a first principles approach to determine eutectic compositions and temperatures. To this end, first, a method to compute the excess molar Gibbs free energy ( $g^{Ez}$ ) of mixing of liquids is described. It combines thermodynamics and statistical mechanics principles applied to alchemical changes conducted using molecular dynamics (MD). Second, by approximating the solid phase properties by those of the pure components, we use the liquid phase  $g^{Ez}$  data to estimate temperature and composition of multicomponent eutectics. To illustrate our approach, we present its application to binary and ternary mixtures of lithium, sodium and potassium nitrate, represented by empirical interatomic potentials. Both calculated eutectic temperature and composition are in good agreement with experimentally determined values[3–6].

**Methodology:** The first step in computing  $g^{Ez}$  as a function of composition ( $x$ ) is to compute  $\Delta\mu$ , the free energy associated with the alchemical transformation of one component into another. In a binary mixture A-B at composition  $x_A$ , for example, an alchemical transformation of particle A into B is performed as a linear function of the transformation variable  $\lambda$ , and the free energy difference for this transformation is computed via

thermodynamic integration[7], in similar vein to Haile’s parameter charging approach[8]. Figure 1 shows this alchemical transformation schematically.

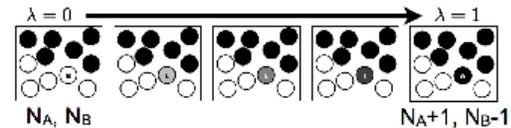


FIG. 1: Alchemical transformation schematic in a binary mixture of A and B. Solid and open particles are of type A and B respectively.

The interaction of the transforming particle with the rest of the mixture is governed by the following potential

$$u_i(\mathbf{r}, \lambda) = \lambda u_{iA}(\mathbf{r}) + (1 - \lambda) u_{iB}(\mathbf{r}), \quad (1)$$

The free energy difference for this transformation is computed from thermodynamic integration as

$$\Delta\mu = \int_0^1 d\lambda \left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda, \quad (2)$$

where  $U$  is the total potential energy of the system,  $U = \sum_i u_i$ . For a binary mixture A-B,  $g^{Ez}$  is computed from  $\Delta\mu$  using

$$g^{Ez}(x_A, x_B) = \int_0^{x_A} dx'_A \Delta\mu - x_A \int_0^1 dx'_A \Delta\mu \quad (3)$$

while for a ternary mixture A-B-C,  $g^{Ez}$  is given by

$$g^{Ez}(x_A, x_B, x_C) = \int_0^{x_A} dx'_A \Delta\mu_{AB} + \frac{-x_A}{x_A + x_B} \int_0^{x_A + x_B} dx'_A \Delta\mu_{AB} + \frac{x_A}{x_A + x_B} g_{AC}^{Ez} + \frac{x_B}{x_A + x_B} g_{BC}^{Ez} \quad (4)$$

where  $\Delta\mu_{AB}$  is the free energy of alchemical transformation of A to B at composition  $(x_A, x_B, x_C)$ .  $g_{AC}^{Ez}$  and  $g_{BC}^{Ez}$  are  $g^{Ez}$  of binary mixtures A-C and B-C respectively.  $x_A$ ,  $x_B$ , and  $x_C$  are component mole fractions, where  $x_A + x_B + x_C = 1$ . Setting  $x_C = 0$  in Eq. 4 yields Eq. 3

For a binary mixture, the free energy of a mixture in a phase is  $g(x_A) = x_A g_A^o + x_B g_B^o + g^{IM} + g^{Ez}$ , where  $g_i^o$  are the free energies of the respective pure components and  $g^{IM} = 1/\beta(x_A \ln x_A + x_B \ln x_B)$ , the ideal free energy of mixing. At phase equilibrium, the chemical

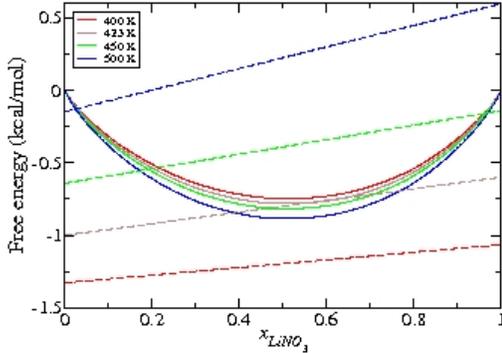


FIG. 2: Demonstration for Li-K system. The solid curves are the free energies of mixing for the liquid phase, while the dashed lines are lines joining the two pure component solid-liquid free energy differences.

potentials of the individual components are equal in the coexisting phases. This requirement can be represented graphically by a common tangent hyperplane drawn between points on the free energy surfaces of the coexisting phases[9]. For solid-liquid phase coexistence, knowledge of the structure and free energy of solid solutions is essential to compute  $g^{\beta x}$ . Although methods have been developed to compute free energy of simple solid solutions [10] and to compute solid-liquid phase equilibria [11–13], for mixtures of more complex molecules, this task of computing free energies of solid mixtures still remains challenging. More complicated, time-consuming efforts have been used to compute phase diagrams of mixtures[14].

The goal of this study is to estimate the eutectic compositions and temperatures, and hence accurate calculation of the entire phase diagram is unnecessary. In our approach, the eutectic temperature and composition are located as the point of tangency of a common tangent hyperplane constructed between pure component solid-liquid free energy differences to the liquid phase free energy of mixing ( $g^{mix} = g^{IM} + g^{\beta x}$ ) surface.  $g^{\beta x}$  is calculated from Eqs. 3, 4 while  $g^{IM}$  is analytic, and the solid-liquid free energy difference is computed using the method outlined in [15]. Since binary and ternary mixtures of lithium, sodium and potassium nitrates are the main focus of this study, henceforth the components of the mixtures will be designated as Li, Na, and K respectively. Figure 2 demonstrates the tangent approach for the Li-K binary mixture.

**Computational details** Simulations were performed on periodic systems of 576 ion pairs at 773.0 K using LAMMPS[16], an open source MD package. The empirical interatomic potential is the same as that used in a previous study of the melting points of pure Li, Na and K nitrates[17]. The three binary systems Li-Na, Li-K

and Na-K, as well as the ternary Li-Na-K system were chosen for this study, and compositions equally spaced over the entire composition range (11 for binary and 45 for ternary) were selected for simulations. The MD simulations were conducted in the NPT ensemble using the Nosé-Hoover thermostat and barostat[18]. Time constants of 0.1 and 0.5 ps were used for the thermostat and barostat, and a timestep of 1 fs was used for the simulations. A cutoff of 12 Å was used for the Buckingham and Coulombic interactions, with tail corrections applied to the Buckingham potential. The particle-particle-mesh (pppm) method[19] was used for long-range electrostatics, with an accuracy of  $10^{-4}$ .

**Eutectic point for mixtures:** The excess free energies in the liquid phase of the binary mixtures LiNO<sub>3</sub>-NaNO<sub>3</sub>, NaNO<sub>3</sub>-KNO<sub>3</sub>, LiNO<sub>3</sub>-KNO<sub>3</sub>, and the ternary mixture LiNO<sub>3</sub>-NaNO<sub>3</sub>-KNO<sub>3</sub> were computed. Free energy differences between liquid and solid phases for pure LiNO<sub>3</sub>, NaNO<sub>3</sub>, and KNO<sub>3</sub> were taken from Ref. [17], a theoretical estimate according to Ref. [15].

Figure 3 illustrates our method for three different temperatures for the ternary mixture. The triangular plane containing the solid-liquid free energy differences for the pure LiNO<sub>3</sub>, NaNO<sub>3</sub> and KNO<sub>3</sub> moves upwards as temperature increases, while the free energy surface for the liquid phase becomes deeper. The temperature at which the triangular plane is a tangent to the liquid free energy surface is when the liquid phase coexists with three solid phases, and hence, is the eutectic. Figure 3(b), the pink arrow maps the location of the eutectic composition onto the triangle diagram.

Binary eutectics for Li-Na and Na-K mixtures were computed from plots similar to figure 2. For ease of locating the ternary eutectic, a fifth order polynomial was fit to  $g^{mix}$  of the liquid phase. Figure 4 compares the computed eutectic lines with experiment. The various binary and ternary eutectics are located on the figure. The polynomial fit captures the curvature of the ternary mixture in the middle, while at the edges which are the location of the binary mixtures, the fit does not perform well. This can be seen in the case of the Na-K mixture. This inaccuracy does not play a role in the estimation of the ternary eutectic. The individual arms of the phase diagram were computed by constructing tangent hyperplanes between the liquid phase free energy and the respective pairs of pure solids corresponding to the low melting binary eutectic being tracked towards the ternary eutectic.

Table I compares the computed eutectic compositions and temperatures for the binary and ternary mixtures against experimental data available in literature. The computed eutectic temperatures and compositions agree well with the experimental data, especially considering no information about solid mixtures was used in computing these eutectic points. We note that the eutectic points computed by assuming ideal mixing for the liquid phase

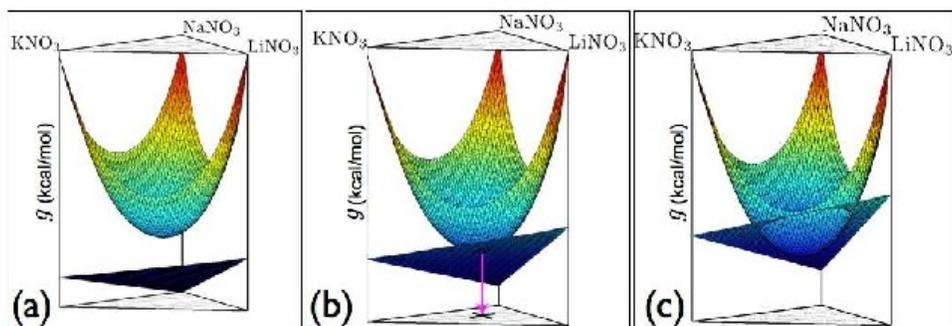


FIG. 3: Computational location of the eutectic point for the ternary mixture Li-Na-K. The figures show free energy surfaces of liquid and solid phases at (a) 398 K, (b) 410 K, and (c) 418 K. The pink vertical arrow in (b) indicates the location of the eutectic composition.

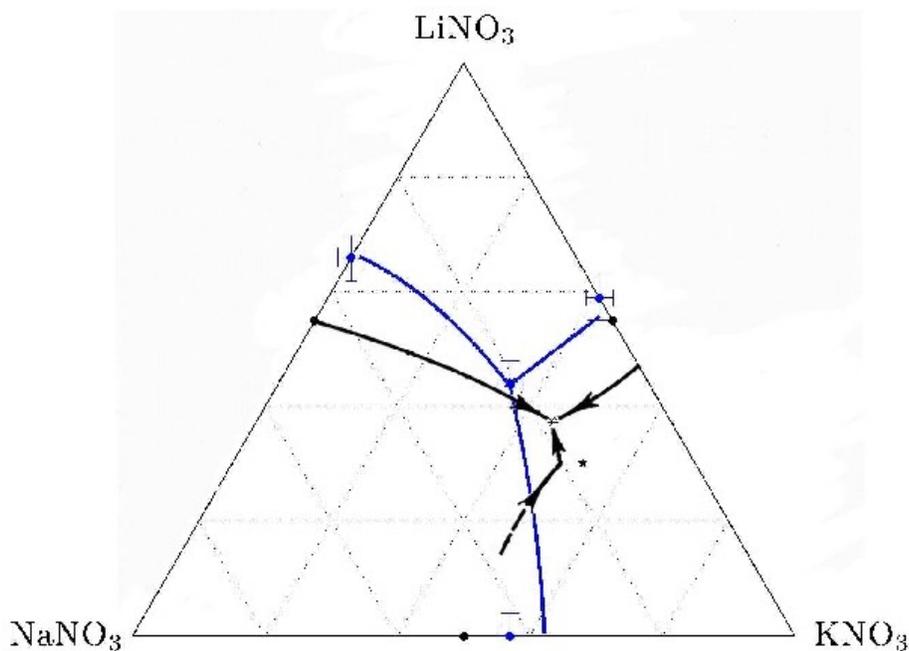


FIG. 4: Comparison of phase diagrams for the Li-Na-K system. Blue symbols and curves denote computed values, while black curves and symbols represent experimental data: black star and curves[3]; black pentagram [4]. The arrows in the black curves denote direction of decreasing temperature. The filled black circles are the experimental eutectic points of the binary mixtures from [5, 6]

TABLE I: Calculated eutectic points, and experimentally determined eutectic points for binary and ternary alkali nitrate mixtures. Compositions are in molefraction units, and the temperatures in K. The uncertainties in the values computed in this study are indicated in brackets

Reference	$x^{(2)}$	$T^{(2)}$	$x_{LiNO_3-MgNO_3}^{(2)}$	$T^{(2)}$	$x_{LiNO_3-KNO_3}^{(2)}$	$T^{(2)}$	$x_{MgNO_3-KNO_3}^{(2)}$	$T^{(2)}$
Bergman and Nogojev[3]	0.375-0.18-0.445	393	-	-	-	-	0.5-0.5	498
Carveth[4]	0.3-0.14-0.56	393	0.5-0.5	479	0.35-0.55	405	0.45-0.55	494
Bergman and Nogojev[20]	-	-	-	-	40.5-59.5	408	-	-
Maeso and Largo [5]	-	-	0.55-0.45	474	0.55-0.45	410	-	-
Kramer and Wilson [6]	-	-	-	-	0.5-0.5	500	-	-
This study (ideal mix)	0.44-0.235-0.325	418	0.66-0.34(0.04)	440(7)	0.59-0.41(0.04)	435(6)	0.42-0.58(0.05)	462(8)
This study (non-ideal mix)	0.44-0.21-0.35(0.04)	411(4)	0.66-0.34(0.04)	435(8)	0.59-0.41(0.04)	423(6)	0.43-0.57(0.05)	460(8)

do not deviate significantly from those for which liquid phase excess free energies were computed using molecular simulations. The compositions show very little deviation, while for the mixtures, the temperature can deviate by more than 10 K. Therefore, for mixtures which do not deviate significantly from ideality, eutectic composition can be estimated using only pure component free energy data. The only computational expense will be in the rigorous calculation of the free energy difference between the solid and liquid phases for the pure components.

*Conclusions:* We used alchemical changes to compute excess free energies of mixtures. The proposed method yields good results for the eutectic composition and temperature. The search space for experimentally locating the lowest melting composition can be reduced drastically using this method. This method is versatile enough to be extended to higher dimensional mixtures provided that good intermolecular potential parameters are available. The free energy difference between the pure solid and liquid phases, and the liquid phase excess molar free energy of mixing are the only information which need to be computed. For mixtures which do not show significant deviation from ideal mixing, mixture calculations can be avoided entirely. Both the alchemical change method and the tangent method can be extended quite easily to higher dimensional mixtures, or other potentials or even *ab-initio* MD. The alchemical method is not limited to empirical potentials, and greater accuracy will be achieved using *ab-initio* methods [21]. Extension of the method to mixtures of monovalent and divalent or multivalent cations will be investigated.

*Acknowledgements:* OAvL acknowledges support from SNL's Laboratory Directed Research & Development Truman program, No. 120209. Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin company, for the U.S. Department of Energy's National Nuclear Security Administration under contract

DE-AC04-94AL85000.

- [1] R. W. Bradshaw and D. E. Meeker, *Sol. Energy Mater.* **21**, 51 (1990).
- [2] R. W. Bradshaw and N. P. Siegel, in *ES2008: Proc. 2nd Int. Conf. Energy Sust., Vol 2* (ASME, New York, 2009), pp. 631-637.
- [3] A. G. Bergman and K. Nogojev, *Zh. Neorg. Khim.* **9**, 1423 (1964).
- [4] H. R. Carveth, *J. Phys. Chem.* **2**, 209 (1898).
- [5] M. J. Maeso and J. Largo, *Thermochim. Acta* **223**, 145 (1993).
- [6] C. M. Kramer and C. J. Wilson, *Thermochim. Acta* **42**, 253 (1980).
- [7] J. G. Kirkwood, *J. Chem. Phys.* **3**, 300 (1935).
- [8] J. M. Haile, *Fluid Phase Equilib.* **20**, 103 (1986).
- [9] C. H. P. Lupis, *Chemical Thermodynamics of Materials* (North Holland, Amsterdam, 1983).
- [10] W. G. T. Kranendonk and D. Frenkel, *Mol. Phys.* **72**, 699 (1991).
- [11] M. R. Hitchcock and C. K. Hall, *J. Chem. Phys.* **110**, 11433 (1999).
- [12] M. H. Lamm and C. K. Hall, *Aiche J.* **47**, 1664 (2001).
- [13] P. A. Apte and I. Kusaka, *J. Chem. Phys.* **123**, 194503 (2006).
- [14] O. Beneš, P. Zeller, M. Salanne, and R. J. M. Konings, *J. Chem. Phys.* **130**, 134716 (2009).
- [15] S. Jayaraman and E. J. Maginn, *J. Chem. Phys.* **127**, 214504 (2007).
- [16] S. J. Plimpton, *J. Comput. Phys.* **117**, 1 (1995), the LAMMPS website is at <http://lammmps.sandia.gov>.
- [17] S. Jayaraman, A. P. Thompson, O. A. von Lilienfeld, and E. J. Maginn, *Ind. Eng. Chem. Res.* **49**, 559 (2010).
- [18] S. Melchionna, G. Ciccotti, and B. L. Holian, *Mol. Phys.* **78**, 533 (1993).
- [19] R. W. Hockney and J. W. Eastwood, eds., *Computer simulation using particles* (McGraw-Hill, 1981).
- [20] A. G. Bergman and K. Nogojev, *Russ. J. Inorg. Chem.* **7**, 179 (1962).
- [21] K. Lemng, S. B. Rempe, and O. A. von Lilienfeld, *J. Chem. Phys.* **130**, 204507 (2009).

**APPENDIX B: QUANTITATIVE STRUCTURE-PROPERTY  
RELATIONSHIPS FOR CHARGE TRANSFER RATES OF  
POLYCYCLIC AROMATIC HYDROCARBONS**

# Quantitative structure-property relationships for charge transfer rates of polycyclic aromatic hydrocarbons

Milind Misra,<sup>1</sup> Denis Andrienko,<sup>2</sup> Björn Baumeier,<sup>2</sup> Jean-Loup Faulon,<sup>3</sup> and O. Anatole von Lilienfeld<sup>1,\*</sup>

<sup>1</sup>*Multiscale Dynamic Materials Modeling Department,*

*Sandia National Laboratories, Albuquerque, New Mexico 87185-1522, USA*

<sup>2</sup>*Max Planck Institute for Polymer Research, Ackermannweg 10, 55128 Mainz, Germany*

<sup>3</sup>*Biology Department, Evry University, Tour Evry2, 523 Terrasses de l'Agora 91034 EVRY cedex*

(Dated: February 11, 2010)

Quantitative structure-property relationships (QSPRs) have been developed and assessed for predicting the reorganization energy  $\lambda$  of polycyclic aromatic hydrocarbons (PAH). Preliminary QSPR models, based on a combination of molecular signature and electronic eigenvalue difference descriptors, have been trained using nearly 200 PAH. Monte Carlo cross-validation systematically improves the performance of the models through progressive reduction of the training set and selection of best performing training subsets. The final biased QSPR model yields correlation coefficients  $q^2$  and  $r^2$  of 0.7 and 0.8, respectively, and an estimated error in predicting  $\lambda$  of  $\pm 0.014$  eV.

## I. INTRODUCTION

A key property of organic semiconducting materials is that their conducting properties can be tuned by optimizing their chemical structure<sup>1-5</sup>. A practical route to do this includes synthesis of a new compound, optimization of its processing conditions, fabrication of the device, and measurement of its performance (properties). By repeating this procedure one can formulate structure-processing-property relationships and proceed with the *rational design* of organic semiconductors.

It is of course tempting to assist this design by first optimizing material properties using computer simulations and modeling. To do this, one has to first devise methods able to predict the property of interest starting from the chemical structure and (preferably) without, or with a minimum of, fitting parameters. The second step consists of correlating these properties with the structure for a specified training set of compounds, and then invert the formulated quantitative structure-property relationship in order to predict optimal compounds for a specific property range.

For organic semiconductors, already the first step in this scheme is non-trivial since charge carrier mobility depends on the electronic structure, local molecular ordering, as well as global percolation pathways for charge carriers. Without going into the details we can say that this is a typical multiscale problem and attempts to solve it constitute an entire research field<sup>6-17</sup>. Our current experience tells us that it is not possible to directly evaluate charge carrier mobility as a property of interest for an *arbitrary* chemical compound, since several assumptions shall be made about the material morphology, type of transport, and the model used to describe it. One could, nevertheless, ask the question whether it is possible to find adequate quantitative structure property relationships (QSPRs) in order to relate chemical structure to charge transport properties, once the link between the structure and mobility is well established.

In this paper, we construct and assess the quality of

several such QSPRs in the context of organic semiconductors. As a test system, we use polycyclic aromatic hydrocarbons (PAHs). PAHs or, more specifically, discotic liquid crystals when condensed, have already found application in organic solar cells and field effect transistors<sup>2,18,19</sup>. A typical chemical structure of a discotic liquid crystal consists of a flat conjugated core with side chains attached to its periphery. Discotics self-assemble into columnar structures with aromatic cores stacked on top of each other. Overlap of the  $\pi$ -orbitals of these cores enables charge transport along columns, rendering these materials one-dimensional semiconductors. The efficiency of charge transport can be engineered by either varying the shape and size of the conjugated core or by influencing their packing by modifying the attached side chains.

Due to structural, dynamical, and energetic disorder, charge transport in discotic liquid crystals occurs via charge carrier hopping between the neighboring molecules. The rate of hopping is given by the high-temperature non-adiabatic Marcus theory<sup>6,20,21</sup>

$$w_{ij} = \frac{J_{ij}^2}{\hbar} \sqrt{\frac{\pi}{\lambda k_B T}} \exp \left[ -\frac{(\Delta G_{ij} - \lambda)^2}{4k_B T \lambda} \right], \quad (1)$$

where  $J_{ij}$  is the electronic coupling matrix element between the neighboring molecules  $i$  and  $j$ ,  $\lambda$  is the reorganization energy,  $\Delta G_{ij}$  is the free energy difference between the initial and final states,  $k_B$  is Boltzmann's constant, and  $T$  is the temperature.

Eq. (1) identifies several important for charge transport parameters. The transfer integral,  $J_{ij}$ , is related to the overlap of electronic orbitals (highest occupied molecular orbital (HOMO) for the hole and lowest molecular orbital (LUMO) for the electron transport). As such, it is very sensitive to the relative position and orientation of neighboring molecules<sup>13,17,22</sup>. In columnar phases of discotics, the maximum of the transfer integral is achieved in a face-to-face molecular arrangement, with the typical intermolecular distance of  $d = 3.5$  Å<sup>19,23-25</sup>. In what follows we assume such an "ideal" molecular arrangement,

since it maximizes charge transport and hence provides an upper bound for the charge mobility which can be reached experimentally. We ignore the distribution in transfer intergals due to thermal fluctuations as well as static defects in morphology.

Another parameter,  $\Delta G_{ij}$ , is the free energy difference between the states with charge localized on the molecule  $i$  or  $j$ . For an ideal face-to-face arrangement this contribution vanishes due to equivalence of the initial and final states.

Finally, an important ingredient influencing charge transport is the internal reorganization energy,  $\lambda$ . It only depends on the chemical structure of a compound, not on processing or morphology. It expresses the strength of electron phonon coupling and has an exponential impact on the transfer rate, with small  $\lambda$  favoring more efficient charge transport.

For an ideal face-to-face columnar alignment the mobility of charge carrier along the column is proportional to the hopping rate with  $\Delta G_{ij} = 0$

$$\mu = \frac{\omega_{ij} d^2}{k_B T} = \frac{J_{ij}^2 d^2}{\hbar k_B T} \sqrt{\frac{\pi}{k_B T \lambda}} \exp\left[-\frac{\lambda}{4k_B T}\right], \quad (2)$$

where  $d$  is the distance between neighboring sites and all other symbols have the same meaning as in Eq. 1. We can therefore argue that large hopping rates (that is large transfer intergals, small reorganization energies) favour high charge mobilities. Hence, the potential descriptors shall link the chemical structure of a compound with the hopping rate, or, alternatively,  $J_{ij}$  and  $\lambda$  values.

In this study we develop appropriate structure-mobility QSPRs. To do this, we first present how the PAH compound data set was generated and used to select the parameters dominating the charge transport in columnar phases of discotics. We then present two descriptors and assess their performance within preliminary QSPR models. Finally, a robust QSPR model is developed using Monte Carlo cross-validation for variable training test-set ratios.

## II. COMPOUND DATA SET

When attempting to predict the charge carrier mobilities of PAHs using yet-to-be specified molecular descriptors, it is not clear *a priori* which of the three physical parameters introduced above dominates the charge transfer rates. In order to (a) facilitate identification of the dominant physical parameter and (b) setup the reference values for our prediction, we have generated a compound data set of PAHs and analyzed its properties. Starting from benzene, we have appended additional aromatic rings at random available bonds. We have used standard carbon-carbon and carbon-hydrogen bond-lengths and angles, checking for atom overlaps as well as aromaticity, and discarding multiple copies of the same PAH. Thus, a data set of 211 closed shell aromatic PAHs with up to nine benzene rings has been generated.

For the remainder of this paper the focus will lie on hole transport. In this case, the reorganization energy can be written as a sum of the relaxation energies in neutral and positively charged state

$$\lambda = E_n^+ - E_n^0 + E_c^0 - E_c^+. \quad (3)$$

Here  $E_g^q$  is an energy of the compound in charge state  $q$  and geometry  $g$ .  $q = 0$  corresponds to a neutral molecule,  $q = +$  to a cation.  $g = n$  indicates optimized geometry of a neutral molecule, while  $g = c$  corresponds to an optimized cation geometry.

According to Eq. 3, the reorganization energy can be easily evaluated using electronic structure calculations by computing four energies on the two potential energy surfaces of the neutral and cationic species.  $\lambda$  has been computed for each compound with density functional theory (DFT) using the B3LYP hybrid-functional<sup>26</sup> and the 6-311++g(d,p) basis set, using the Gaussian 03<sup>27</sup> program suite. Four calculations per compound are necessary, geometry optimizations for the neutral ( $E_n^0$ ) and cationic ( $E_c^+$ ) species, and single point energy calculations for the cationic species in the neutral geometry ( $E_n^+$ ) as well as for the neutral species in the cationic geometry ( $E_c^0$ ). Fig. 1(a) displays the values of  $\lambda$  in ascending order together with their distribution in the data set. The 211 compounds have an average reorganization energy of 0.13 eV, with a spread from 0.06 to 0.30 eV (see also Table III for more details).

As was mentioned above, the transfer integrals  $J_{ij}$  in Eq. (1) depend on the relative geometrical configuration of two molecules. For our set of PAHs, these transfer integrals have their maximum value for a *face-to-face* orientation, in which the overlap of the  $\pi$ -orbital system is maximal. Typically the molecules arrange at a van der Waals distance of about 3.5 Å. Therefore, we have calculated  $J_{ij}$  for such a co-facial geometry (or columnar alignment) at such a mutual distance, based on Zerner's Intermediate Neglect of Differential Overlap (ZINDO) using the Molecular Orbital Overlap (MOO) package<sup>28</sup>.

Panel (b) of Fig. 1 shows the resulting values of  $J$  in ascending order and its distribution within the data set. The transfer integrals span a range of 0.1 to 0.5 eV, which is relatively large due to the assumed columnar stacking of the molecules. The distribution is rather sharply peaked at about 0.4 eV, indicating that there are only faint variations of  $J$  within the data set. Using the values obtained for the reorganization energy and the transfer integral, we have evaluated Eq. (1) to determine the transfer rates  $\omega_{ij}$  for each compound. The corresponding distribution is shown in Fig. 1(c).

Finally, Fig. 1(d) combines the three parameters  $x = \lambda, J, \omega$  plotted as a function of the  $\lambda$ -sorted compound index. To facilitate an easy comparison, all values are shown relative to the value of the compound with index zero ( $x_0$ ). This representation illustrates that among the three parameters the reorganization energy has the largest relative variance, and that the transfer integrals only slightly fluctuate around a constant value. The lat-

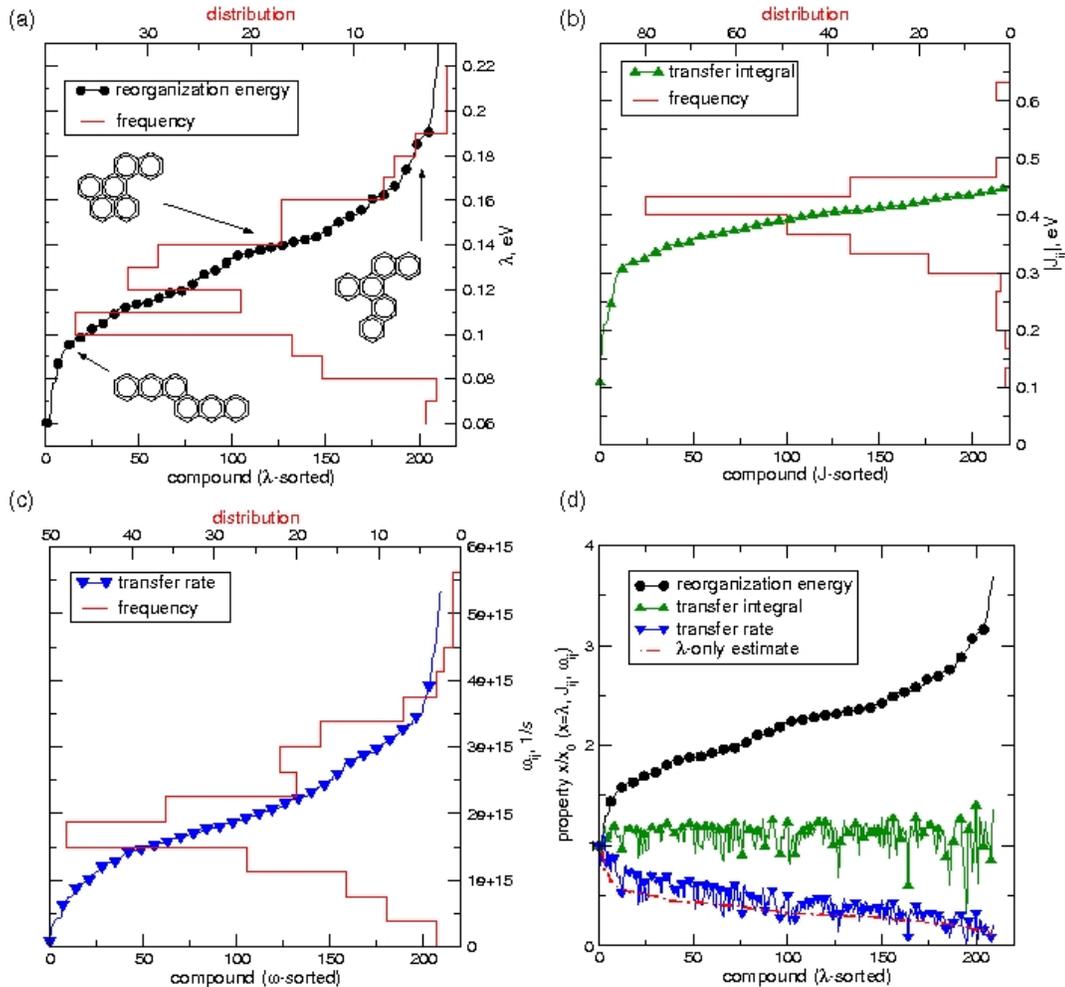


FIG. 1: (Color online) Analysis of reorganization energies, transfer integrals, and transfer rates in the PAH compound data set: Sorted values of (a)  $\lambda$  [black circles], (b)  $J$  [green up triangles], and (c)  $\omega$  [blue down triangles] as functions of compound indices, as well as the respective compound distributions (red lines). For illustrative purposes, three compounds at  $\lambda$  values 0.09, 0.14, and 0.18 eV are shown as insets in panel (a). Panel (d) shows a combined plot of the three properties against the  $\lambda$ -sorted compound index. The red dash-dotted line indicates a  $\lambda$ -only prediction of the transfer rates based on Eq. (1) assuming a uniform transfer integral for the data set.

ter is a good indication that it is indeed mainly the reorganization energy that dominates the transfer rates, and thereby the charge carrier mobilities via Eq. (2). To emphasize this notion, we have also included a  $\lambda$ -only estimate of the transfer rates, which is based on the assumption of a constant transfer integral for the entire data set. The result (shown in Fig. 1(d) by the red dash-dotted line) corroborates the notion of  $\lambda$ -dominated charge transfer. Based on this analysis one can conclude

that a prediction of charge transfer properties/mobilities in PAHs from molecular descriptors can be reduced to the prediction of the molecular reorganization energy. In the following section we will analyze the performance of different descriptors referring to results presented in this section.

### III. RESULTS AND DISCUSSION

Any attempt to develop analytical but quantitative statistical relationships between physical properties of a compound and its structure relies on the definition of variables (descriptors) that characterize the underlying atomistic structure. Many descriptors are conceivable including, for instance, those that are exclusively based on geometrical molecular features such as deviations from planarity or linearity. Appendix A describes several of such scalar descriptors that we have investigated but rejected simply due to their low correlation with the reorganization energy (see Table III). For the sake of a compact presentation, we limit the discussion in this section to the two descriptors (*signature* and  $\Delta\epsilon$ ) that turned out to have the largest correlation with  $\lambda$ . We first discuss their use for the preliminary QSPR models that are based on the full 211-compound PAH data set. Then we present the development of “biased” QSPR models after partitioning the data set into a 188-compound training set and a 23-compound test set. Finally, we discuss the validity of the descriptors in the light of their predictive power regarding reorganization energies of the test set compounds.

#### A. Molecular signature

The molecular *signature* is a compilation of a set of atomic *signatures*,  $\{\sigma\}$  that occur in a molecule. It was first presented and applied in the context of structure elucidation<sup>29</sup>, and later defined for acyclic compounds and used in QSPR analyses<sup>30</sup>. An atomic *signature* describes the extended covalent bonding neighborhood of an atom within a molecule up to a certain “height” ( $h$ ). See Fig. 2 for more details on how atomic *signatures* ( ${}^h\sigma$ ) are generated. The molecular *signature* for a given height is a vector that contains the frequencies of all the  ${}^h\sigma$  occurring in the molecule. As such it represents a methodical codification system over an alphabet of atom types.

The MolConverter program from ChemAxon<sup>31</sup> was used to convert the *xyz*-coordinate files of the structures in our PAH data set to corresponding SMILES (simplified molecular input line entry specification) strings. SMILES describe chemical structures and topologies using short textual strings<sup>32</sup>. From the SMILES strings molecular *signatures* have been determined for individual heights. The straight-forward correlation between these molecular *signatures* and the reorganization energy is insufficient for predictive quality. For example, the correlation coefficient,  $r^2$ , of molecular *signatures* with  $h \in \{0, 1, 2, 3\}$  versus  $\lambda$  does not exceed 0.39.

#### B. HOMO eigenvalue difference, $\Delta\epsilon$

As mentioned above, the reorganization energy  $\lambda$  expresses the strength of electron-phonon coupling in the

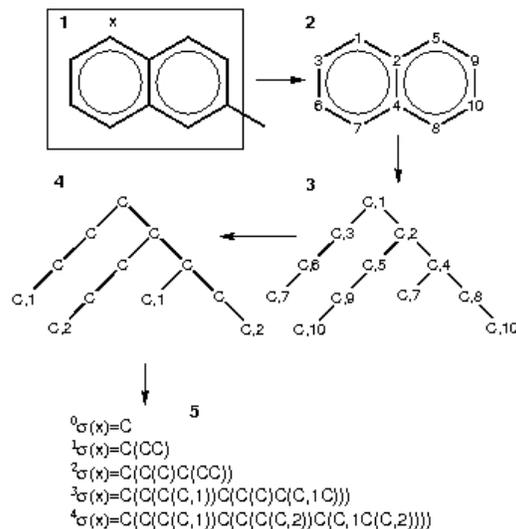


FIG. 2: The figure shows atomic *signatures* ( ${}^h\sigma$ ) from height  $h = 0$  to 4 for an exemplary atom X in 2-methyldecalhydronaphthalene.  ${}^h\sigma$  of atom X is determined as follows, (1) The subgraph containing all atoms at distance 4 from atom X is extracted. (2) This subgraph is canonicalized with atom X having label 1. (3) A tree spanning all edges of the subgraph is constructed. (4) All labels appearing only once are removed and the remaining labels are renumbered in the order they appear. (5) The atomic *signature* is determined after reading the tree in a depth-first order, the depth corresponding to height  $h$ .

molecule. Thus, it is not surprising that a descriptor based solely on structural features, such as the molecular *signature*, correlates with  $\lambda$  only insufficiently. We have therefore leveraged the notion of variable number of electrons,  $N_e$ , such as put forth within conceptual DFT<sup>33</sup>, in order to associate *electronic* properties to descriptors that improve this correlation.

First, we note that Eq. (3) can be rearranged in terms of the difference between vertical “excitation” energies,  $\delta$ , linking states of same geometry but different  $N_e$ , yielding

$$\lambda = [E_n^+(N_e - 1) - E_n^0(N_e)] - [E_c^+(N_e - 1) - E_c^0(N_e)] = \delta_n - \delta_c. \quad (4)$$

Here,  $\delta_n$  is the iso-nuclear change in energy due to removal of an electron from the neutral species in its relaxed geometry, while  $\delta_c$  is the iso-nuclear change in energy due to addition of an electron to the cationic species in its relaxed geometry.

Referring to the molecular grand-canonical ensemble DFT-based exploration of chemical compound space<sup>34–36</sup>, we can estimate the  $\delta$  contributions in Eq. (4) through first order Taylor expansions in number of elec-

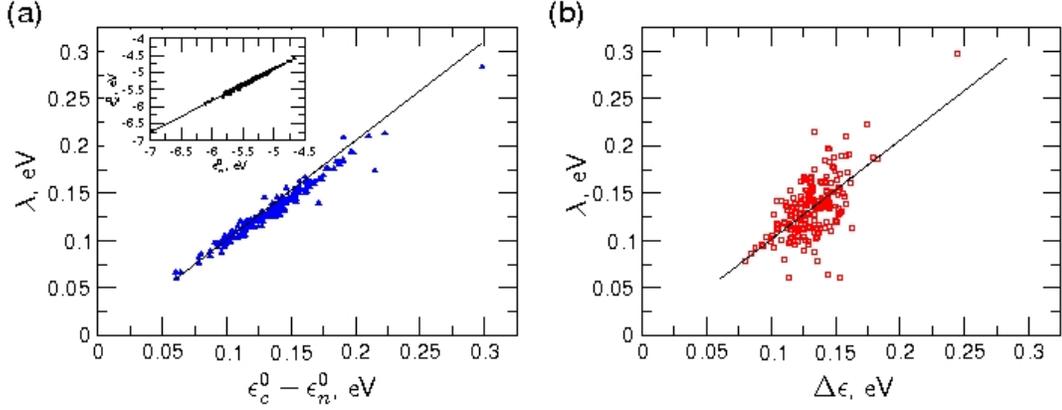


FIG. 3: (Color online) Correlations of calculated reorganization energies  $\lambda$  of the PAH data set with differently predicted values. In panel (a), the correlation to  $\epsilon_c^0 - \epsilon_n^0$  according to Eq. (4) results in  $\lambda = 1.05 \times (\epsilon_c^0 - \epsilon_n^0) - 0.004$  [eV] with  $r^2 = 0.96$  (blue triangles). The inset shows the correlation between the highest occupied electronic eigenvalue of the neutral molecule in cationic geometry,  $\epsilon_c^0$ , with the respective value for the neutral geometry  $\epsilon_n^0$ , resulting in  $\epsilon_c^0 = 0.93 \times \epsilon_n^0 - 0.25$  [eV] and  $r^2 = 0.99$ . Using this relation to predict  $\lambda$  effectively from  $\epsilon_n^0$  only according to  $\Delta\epsilon$  in Eq. (6) yields the correlation shown in panel (b) with  $\lambda = 1.01 \times \Delta\epsilon + 0.001$  [eV],  $r^2 = 0.39$  (red squares).

trons,  $N_e$ . Specifically they read

$$E(N_e + \Delta N_e) = E(N_e) + \frac{\partial E(N_e)}{\partial N_e} \Delta N_e + \mathcal{O}(\Delta N_e^2). \quad (5)$$

For an exact expression for the exchange-correlation potential within density-functional theory, all higher order terms would vanish for  $0 \leq \Delta N_e \leq 2$  because the total potential energy of a molecule with fixed external potential changes only linearly as one varies the number of electrons<sup>37-39</sup>. Since the derivative of the energy with respect to  $N_e$  is given as the eigenvalue of the HOMO<sup>40</sup> ( $\epsilon$ ), we can combine Eqs. (4, 5) and express  $\lambda$  as

$$\begin{aligned} \delta_n &= \frac{\partial E_n^0(N_e)}{\partial N_e} \Delta N_e = \epsilon_n^0(N_e) \Delta N_e \\ \delta_c &= \frac{\partial E_c^0(N_e)}{\partial N_e} \Delta N_e = \epsilon_c^0(N_e) \Delta N_e \\ \lambda &= \epsilon_c^0 - \epsilon_n^0, \end{aligned} \quad (6)$$

where  $\Delta N_e = -1$ , and  $\epsilon_n^0(N_e)$  and  $\epsilon_c^0(N_e)$  denote the eigenvalues of the highest occupied molecular Kohn-Sham orbitals of the neutral molecule in the respective optimal neutral and cationic geometry.

However, the exact form of the exchange-correlation functional is not only unknown, it has also been shown that the self-interaction error increases for fractional occupation within widely used functionals<sup>41</sup>. The difference between electronic eigenvalues of the HOMOs in the neutrally and cationically relaxed geometries  $\epsilon_c^0 - \epsilon_n^0$  yields therefore only an estimate of  $\lambda$  due to the use of approximate functionals. In our case, we have tested the quality of this approximation based on the B3LYP hybrid-functional by correlating the difference in  $\lambda$  obtained

from the eigenvalues as in Eq. (6) with  $\lambda$  obtained from the energies according to Eq. (3). As shown in Fig. 3(a), the correlation is very strong with a correlation coefficient  $r^2$  of 0.96. It seems likely that the remaining deviation is indeed due to the approximate nature of B3LYP. The use of a functional that correctly accounts for fractional occupation numbers, such as developed by Mori-Sánchez et al.<sup>39</sup>, can be expected to improve this correlation even further.

The (approximate) determination of the reorganization energy according to Eq. (6) still requires the calculation of three energies on the potential energy surface, i.e. the optimizations of neutral and cationic geometries, as well as a single-point calculation for the neutral molecule in the cationic geometry. While this is one calculation less than in Eq. (3), it is inconvenient since ideally one would like to predict  $\lambda$  from ground-state properties of the neutral molecule alone, i.e., without having to calculate  $\epsilon_c^0$ . We have therefore probed if  $\epsilon_c^0$  correlates with  $\epsilon_n^0$  in the PAH data set. The inset in Fig. 3(a) shows  $\epsilon_c^0$  plotted versus the respective  $\epsilon_n^0$ . One can clearly identify a linear relationship and predict  $\epsilon_c^0$  using  $\epsilon_n^0$  only. The linear regression yields  $\epsilon_c^{\text{pred}} := 0.93 \times \epsilon_n^0 - 0.25$  [eV] with a remarkable correlation of  $r^2 = 0.99$ .

Based on these two relations, we have used  $\Delta\epsilon$  as a scalar descriptor for  $\lambda$  such that

$$\lambda \approx \Delta\epsilon \equiv \epsilon_c^{\text{pred}} - \epsilon_n^0. \quad (7)$$

Fig. 3(b) shows the correlation of the actual  $\lambda$  from Eq. (3) with the estimate  $\Delta\epsilon$ ,  $\lambda \approx 1.01 \times \Delta\epsilon + 0.001$  [eV]. The regression for this expression, however, yields a rather low correlation coefficient of only  $r^2 = 0.39$ .

TABLE I: Preliminary QSPR models, (i) - (viii), and corresponding  $q^2$  values for multiple linear regression (MLR) and partial least squares (PLS), respectively. See Section III C and Appendix A for more details.  $h$  is the *signature* height,  $\#\sigma$  refers to the number of atomic *signatures*, i.e. the dimension of the molecular *signature* vector. These models were generated using the data set of 211 PAHs. Highlighted model (ii) has been used for the construction of the "biased" QSPR.

$h$	type	$\#\sigma$	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
0-3	MLR	63	0.29	0.62	0.20	0.28	0.29	0.29	0.63	0.64
0-3	PLS	63	0.47	0.47	0.47	0.46	0.47	0.47	0.46	0.45
0-4	PLS	431	0.50	0.50	0.50	0.49	0.50	0.50	0.50	0.26
0-5	PLS	1635	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.31

(i): molecular *signatures*  
(ii): molecular *signatures* +  $\Delta\epsilon$   
(iii): molecular *signatures* +  $dM$   
(iv): molecular *signatures* +  $dL$   
(v): molecular *signatures* +  $dP$   
(vi): molecular *signatures* +  $dH$   
(vii): molecular *signatures* +  $dM$  +  $dL$  +  $dP$  +  $dH$  +  $\Delta\epsilon$   
(viii): (vii), redundant descriptors removed based on UFS

While it seems possible to remove any reference to the cationic geometry from the estimate of  $\lambda$  using the electronic eigenvalue descriptor  $\Delta\epsilon$  in Eq. (7) it is also clear that this solely electronic descriptor alone is not sufficient to reliably predict reorganization energies.

### C. Preliminary QSPR models from combining molecular *signatures* with $\Delta\epsilon$

From the two preceding sections, it is apparent that when used separately neither the structural molecular *signature* nor the electronic eigenvalue descriptor  $\Delta\epsilon$  are sufficient for reliable quantitative estimates of the reorganization energy in our set of PAHs. Since  $\lambda$  is a measure of the coupling of structural and electronic degrees of freedom in a molecule, it is natural to attempt a combination of the two descriptors. For the different heights of molecular *signatures* (see Section III A), we have set up different preliminary QSPR models using *signatures* of heights 0-3 through 0-5 for the PAH compound data set (without outliers). Specifically, leave-one-out cross-validated correlation coefficient ( $q^2$ ) have been calculated using multiple linear regression (MLR) and partial least squares (PLS). These coefficients, together with the preliminary models, are listed in Table I, and more technical details are given in Appendix B. For the sake of completeness, we also present the results for additional models that are not based on  $\Delta\epsilon$  but that combine molecular *signature* with various other scalar structural descriptors. These additional models proved less promising and are explained in Appendix A.

Our results show that while the PLS calculations yield a  $q^2$  of around 0.50, indicating predictability in general for all the models, they do not suggest a preference for a

particular descriptor combination. MLR results, in contrast, indicate a clear preference for the model combination of molecular *signature* with  $\Delta\epsilon$  [model (ii) in Table I], which has a  $q^2$  of 0.62. For the alternative combinations of molecular *signature* and various structural scalar descriptors the corresponding  $r^2$  ranges only from 0.20 to 0.29. An additional model which combines all descriptors considered in this study [model (vii) in Table I] does not improve the performance of preliminary QSPR model (ii), even when using unsupervised forward selection of descriptors (see Appendix B 2) to further eliminate redundancy among descriptors [model (viii)].

Thus, the combination of height 0-3 *signatures* and  $\Delta\epsilon$  in model (ii) is identified as the optimal starting combination for developing the "biased" QSPR model in the next section.

### D. Biased QSPR models from Monte Carlo cross-validation

We found that the previously identified optimal preliminary QSPR model can further be developed into a "biased" QSPR model with the help of Monte Carlo cross-validation. To this end, the PAH compound data set was first split into a total 188-compound training set, and a test set of 23 compounds, enabling the *a posteriori* validation of the biased models. The test set was determined using dissimilarity based compound selection, as described in Appendix B 3, and the resulting test set compounds are shown in Table II.

Thereafter, out of the total 188-compound training set subsets with varying percentages  $x$  were defined, where  $x \in (5, 10, 15, \dots, 90, 95)\%$ . For each  $x$ , 10,000 random partitions from among the 188 compounds were generated. All the random partitions were subjected to training using the preliminary QSPR model (ii), i.e. height 0-3 molecular *signatures* combined with  $\Delta\epsilon$  based on PLS.

The models obtained, dubbed  $M_x^k$  ( $k \in 1, 2, \dots, 10,000$ ), were subsequently ranked according to their performance as measured by  $q^2$ . For  $M_{100}$ ,  $q^2 = 0.44$  and  $r^2 = 0.53$ , we note that  $q^2(M_{100})$  lies below the conventional predictive threshold of 0.50. As described in more detail in Appendix B 4,  $q^2$  can be improved by reducing the training subset size  $x$ , followed by the QSPR model training of 10,000 random partitions for each of these reduced training subsets.

Fig. 4 illustrates the results for varying percentage  $x$ . The average  $q^2$ , i.e. the average of the cross-validated correlation coefficients over all randomly chosen partitions, declines progressively as the training set size decreases. The standard deviation around that average, however, increases even more, thereby enabling us to identify "biased" models  $M_x^k$ , namely models that yields the best  $q_x^2$  out of all the 10,000 models that have been trained for each partition at a particular  $x$ . This behavior is in line with Ref. 42.

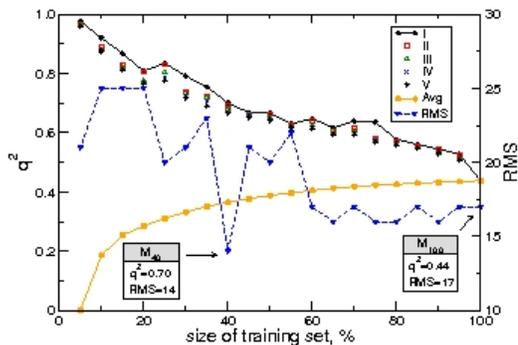


FIG. 4:  $q^2(M_x^k)$  of the best five partitions  $k \in \{I, II, III, IV, V\}$  (out of 10,000) as a function of training subset size  $x$  in steps of 5%. Yellow represents the average  $q^2$  of 10,000 partitions. Root mean square (RMS) deviation [meV] of actual  $\lambda$  from  $\lambda$  predicted by  $M_x^{k=I}$  for test set compounds in Table II.  $M_{40}$  refers to the best partition (I), and is dubbed the biased QSPR model.

#### E. Test set results of biased QSPR models

Fig. 4 suggests that making predictions of the reorganization energy based on the biased (best performing) Monte-Carlo models will always be the more favorable the smaller the subset. From the behavior of  $q^2$  versus  $x$  one could be tempted to deduce that the optimal model should be based on the smallest training subset. However, it is clear that this is not a sufficient requirement for the predictability of a model, and that external (*a posteriori*) validation is a sounder way to assess the reliability of a QSPR model<sup>43</sup>. Thus, to determine the optimal size of the training subset, we have computed the root mean square (RMS) deviation of predicted  $\lambda$  from actual  $\lambda$  for the 23 test set molecules using the biased  $M_x^I$ , where  $x = 5, 10, \dots, 100$  (Note that we excluded two outliers from the test set since they had the largest residuals and corresponded to extreme  $\lambda$  values (maximum and minimum) within the entire compound data set).

As shown in Fig. 4, as  $x$  decreases from 100 to 60 % RMS remains roughly constant ( $\sim 16$  meV), and starts to strongly increase in oscillatory fashion for subsets smaller than 55 %. Since RMS is minimal at  $x = 40\%$  (14 meV), we define the corresponding biased model  $M_{40}$  as our best QSPR model for predicting  $\lambda$  of PAHs. In contrast, model  $M_{100}$  has a higher RMS of 17 meV. Biased QSPR model  $M_{40}$  does not only have a lower RMS deviation but also exhibits improved correlation coefficients,  $q^2 = 0.70$  and  $r^2 = 0.80$ . Table II lists the residuals for the predictions of  $\lambda$  based on models  $M_{100}$  and  $M_{40}$ .

In summary, model  $M_{100}$  predicts the reorganization energy of more than 75% compounds within a reasonable margin of error ( $\pm 20$  meV). Biased model  $M_{40}$ , however, predicts a larger number ( $>85\%$ ) of test set compounds

within the same error margin  $\pm 20$  meV.

## IV. CONCLUSIONS

Based on conceptual density functional theory, we have developed a new frontier orbital eigenvalue descriptor  $\Delta\epsilon$  for the empirical prediction of reorganization energies,  $\lambda$ . For a compound data base of over 200 polycyclic aromatic hydrocarbons, we have investigated the suitability of a combination of a structural (molecular *signature*) and an electronic ( $\Delta\epsilon$ ) descriptor for quantitative structure property relationship (QSPR) models of reorganization energies. For the entire data set, we find that preliminary QSPR models yield at best a correlation coefficient of  $q^2 = 0.5$ . Monte Carlo cross-validation with training subsets of reduced sizes enabled us to identify a “biased” model,  $M_{40}$ , with markedly better performance. Specifically,  $M_{40}$  yields a  $q^2$  and  $r^2$  of 0.70 and 0.80, respectively, and a root mean square deviation of predicted from actual  $\lambda$  of only 0.014 eV. Additional scalar structural descriptors, such as average interatomic distance, deviation from linearity, or deviation from planarity have been devised and tested, but yielded only negligible improvement when combined with molecular *signature*. We can also support the basic assumption of selection algorithms based on dissimilarity which requires that compounds spanning structure/descriptor space also span property/activity space.

We conclude that a biased QSPR, based on a combination molecular *signature* and  $\Delta\epsilon$  and Monte Carlo cross-validated training, appears sufficiently reliable to be used to for the routine prediction of upper bounds of charge carrier mobilities in discotic liquid crystals.

## V. SUPPLEMENTAL MATERIALS

PLS parameters for the robust models  $M_{40}$  and  $M_{100}$ .

## VI. ACKNOWLEDGMENTS

MM and OAvL acknowledge support from SNL Truman Program LDRD project No. 120209. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy’s National Nuclear Security Administration under contract DE-AC04-94AL85000. This work was partially supported by DFG via IRTG program between Germany and Korea, DFG grants AN 680/1-1 and SPP1355. D.A. acknowledges the Multiscale Materials Modeling Initiative of the Max Planck Society.

TABLE II: Actual  $\lambda$ 's [eV] and residuals [meV] of predicted  $\lambda$ 's of test set of compounds for the two biased models,  $M_{100}$  and  $M_{40}$ .

ID	$\lambda$	$M_{100}$	$M_{40}$	structure	ID	$\lambda$	$M_{100}$	$M_{40}$	structure	ID	$\lambda$	$M_{100}$	$M_{40}$	structure
0	0.30	-111	-100		1	0.19	1	8		2	0.22	-43	-27	
7	0.11	4	5		41	0.13	-5	5		43	0.13	-9	-10	
55	0.20	-21	-25		83	0.11	8	13		86	0.14	-10	-15	
107	0.11	11	-11		116	0.10	7	2		124	0.12	-3	6	
129	0.09	-11	-9		143	0.13	6	8		158	0.12	-2	2	
166	0.11	13	13		185	0.13	-13	-4		193	0.10	6	8	
195	0.09	25	18		207	0.06	72	77		209	0.10	31	20	
211	0.16	-34	-29		215	0.15	-3	0						

## Appendix A: Descriptors

### 1. Scalar descriptors

In addition to *signatures* for various heights and  $\Delta\epsilon$ , four additional scalar descriptors have been considered, which will be described in the following.

#### a. Molecular average distance, $dM$

$dM$  is defined as the average distance in a molecule calculated from Cartesian atomic coordinates,  $\{r_i\}$

$$dM = \frac{1}{N} \sum_{i < j} |r_i - r_j|$$

where  $N$  is the total number of atom pairs,  $i$  and  $j$ .  $dM$  is an index of the spatial extension of a molecule. Since usually larger molecules accommodate charges more easily one would expect  $dM$  to correlate with smaller values of  $\lambda$ . It has therefore been included in the list of possible descriptors.

#### b. Deviation from planarity, $dP$

$dP$  expresses the 3-dimensional molecular curvature after geometry optimization of the neutral molecule. This descriptor is meant to represent the perfection in the delocalized  $\pi$ -electron system of the molecule. One would expect that the more the system is delocalized, the more easily it accommodates a positive charge, and the smaller its  $\lambda$  is.

An orthogonal regression that minimized the perpendicular distances from the atomic coordinates to a fitted molecular plane was carried out using principal components analysis. Thus, while the coefficients of the first two principal components define the basis set vectors for the fitted plane, the third principal component provides the coefficients for the normal to the plane. For each molecule, the average sum of squared deviations of all atoms from the plane was used as the descriptor ( $dP$ ) in the QSPR calculations

$$dP = \frac{1}{N} \sum_i |r_i - r_i^P|^2, \quad (A1)$$

where  $r_i$  is the vector defining the position of atom  $i$ ,  $r_i^P$

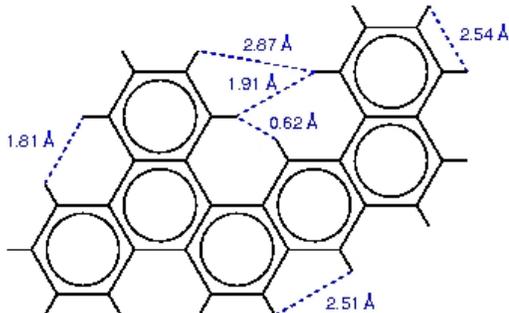


FIG. 5: Exemplary structure displaying all the various types of generated H-H default distances:  $dH_1 = 0.62 \text{ \AA}$ ,  $dH_2 = 1.81 \text{ \AA}$ ,  $dH_3 = 1.91 \text{ \AA}$ ,  $dH_4 = 2.51 \text{ \AA}$ ,  $dH_5 = 2.54 \text{ \AA}$ , and  $dH_6 = 2.87 \text{ \AA}$ .

is the projection of the atom on the fitted plane, and  $N$  is the total number of atoms in the molecule.

### c. Deviation from linearity, $dL$

$dL$  represents deviation from linearity.  $dL$  expresses the 2-dimensional molecular curvature of the generated PAHs. In this case, two principal components were required; the first component defines the vector for the fitted line through Cartesian coordinates, and the second component provides coefficients for the perpendicular to the fitted line. This descriptor,  $dL$ , was defined as the average sum of squared deviations of all atoms from the line

$$dL = \frac{1}{N} \sum_i |\mathbf{r}_i - \mathbf{r}_i^L|^2, \quad (\text{A2})$$

where  $\mathbf{r}_i^L$  is the projection of atom  $i$  on the fitted line. These calculations (and the calculation of  $dP$ ) were performed using the *princomp* method in MATLAB<sup>44</sup>.

### d. Number of close hydrogens, $dH$

$dH$  enumerates close hydrogen atoms and is related to the compound's likelihood to deviate from planarity. A given PAH may contain steric hydrogen-hydrogen repulsions that force it to assume a non-planar geometry. The number of such interactions in each molecule has been included as another descriptor (hydrogen interaction descriptor or  $dH$ ). We considered possible H-H interactions (up to  $3.0 \text{ \AA}$  distances) in the PAHs with the default geometry, i.e., with structures generated by the growth algorithm, using default bond lengths, as displayed in Fig. 5.

Linear regression of the  $dH_1$  vector, consisting of presence (1, 2, or 3 in number) or absence (0) of one or more

H-H distances of  $0.62 \text{ \AA}$  with  $dP$  results in a correlation coefficient of 0.80 confirming a strong positive correlation with deviation from planarity. Hence, the  $dH_1$  H-H interaction vector was included as the hydrogen interaction descriptor,  $dH$ . Since its calculation does not require geometry optimization, it would eventually be desirable to replace  $dP$  by  $dH$ .

## 2. Performance of scalar descriptors

Table III lists descriptive statistics for  $\lambda$  and the scalar descriptors. The last row in the table lists the correlation coefficient of the descriptors with  $\lambda$ . The highest occupied molecular orbital eigenvalue based and molecular distance based descriptors correlate with  $\lambda$  somewhat better than all others, suggesting absence of a linear relationship between the latter and  $\lambda$ .

TABLE III: Descriptive statistics for reorganization energy  $\lambda$  and all the scalar descriptors defined in Section A 1. The descriptors correspond to molecular average distance ( $dM$ ), deviation from linearity ( $dL$ ), deviation from planarity ( $dP$ ), number of close hydrogens ( $dH$ ), and  $\Delta\epsilon$ .

	$\lambda$ [eV]	$dM$ [Å]	$dL$ [Å <sup>2</sup> ]	$dP$ [Å <sup>2</sup> ]	$dH$ [Å]	$\Delta\epsilon$ [eV]
Mean	0.13	5.85	5.17	0.14	0.52	0.13
Std Err	0.00	0.04	0.10	0.01	0.05	0.00
Median	0.14	5.85	5.11	0.00	0.00	0.13
Std Dev	0.03	0.59	1.48	0.22	0.66	0.02
Kurtosis	3.68	4.00	0.88	7.16	0.86	5.76
Skewness	0.92	-1.09	0.65	2.25	1.10	0.88
Minimum	0.06	2.77	2.02	0.00	0.00	0.08
Maximum	0.30	7.09	11.54	1.33	3.00	0.25
$r^2$	-	0.27	0.00	0.07	0.07	0.39

## 3. Details of the molecular signatures

The number of unique atomic *signatures* calculated for height 0 through height 5 are listed in Table IV. For instance, since the height 0 atomic *signature* is simply the atom type itself, there are only two unique atomic *signatures* ([C] and [H]) corresponding to these two atom types in any given PAH. The collection of atomic *signatures* for heights 0-3 (63 *signatures*) was found to be sufficient to produce a descriptor matrix with the minimal number of atomic *signatures* required to uniquely represent every molecule in the data set. Thus, whereas heights 0-1 and 0-2 produced an insufficient number of atomic *signatures* (5 and 12, respectively) for this purpose, heights 0-4 and 0-5 (431 and 1635 atomic *signatures*, respectively) resulted in over description of molecules in the data set. It turns out (see Table I) that going beyond height 3 fails to improve the QSPR models.

TABLE IV: Number of atomic signatures for increasing heights (for 213 PAH's)

Height	0	1	2	3	4	5
Total	426	638	1368	3352	5222	6634
Unique	2	3	7	51	368	1206

## Appendix B: Methodology of the QSPR analysis

### 1. Statistical evaluation

Unlike MLR, PLS is capable of handling short and wide (or over-square) matrices such as those frequently encountered in QSAR/QSPR studies<sup>45,46</sup>. In this study, the descriptor matrix consisted of rows corresponding to compounds and columns corresponding to descriptors. PLS and MLR with leave-one-out cross-validation calculations were performed in MATLAB using the *plsregress* and *regress* functions, respectively. Whereas PLS was used for all combinations of heights and descriptors, MLR was used only when the input matrix had appropriate dimensions. For all models, the correlation coefficient ( $r^2$ ) and the leave-one-out cross-validated correlation coefficient ( $q^2$ ) were calculated by

$$r^2 = 1 - \frac{SSE}{SST}$$

$$q^2 = 1 - \frac{PRESS}{SST}$$

where *SSE* is sum of squared errors, *SST* is the total sum of squares corrected for the mean, and *PRESS* is the prediction error sum of squares. For PLS models the optimum number of components, *c*, was also determined with *c* = 6 taken as the cutoff for maintaining a low complexity model.

### 2. Unsupervised forward selection (UFS)

For some initial models, redundant descriptors were removed resulting in a maximally orthogonal subset. Program of Whitley *et al.*<sup>47</sup> was used for data reduction. Briefly, the UFS algorithm<sup>48</sup> first selects the two descriptors with the smallest pairwise correlation coefficient. It then rejects all descriptors whose pairwise correlation coefficient with the first two descriptors exceeds a user-specified value,  $r_{max}^2 < 1$ . The algorithm continues iteratively until all descriptors are either selected or rejected. A descriptor is selected if it has the smallest squared multiple correlation coefficient (*smcc*) with previously selected descriptors and all descriptors with  $smcc > r_{max}^2$  with currently selected descriptors are rejected. In this study,  $r_{max}^2$  was set to 0.99. In addition, descriptors with a standard deviation near zero ( $< 0.005$ ) were also removed.

### 3. Test set generation

Given *n* data points, suppose they are split in two parts. The first part, containing  $n_{tr}$  data points, is used for training a statistical model. The second part, called the test set, contains  $n_{te} = n - n_{tr}$  data points and is set aside for model validation, i.e., for evaluating the predictive ability of the model trained on  $n_{tr}$  data points. Dissimilarity based compound selection was used to select  $n_{te}$  test set compounds as follows. First, the Tanimoto coefficient<sup>49</sup>, *S*, was used to compute the pairwise similarity between compounds in the data set:

$$S(M_i, M_j) = \sum_{k=1}^m \frac{d_{ik}d_{jk}}{d_{ik}^2 + d_{jk}^2 - d_{ik}d_{jk}},$$

where  $M_i$  and  $M_j$  are the  $i^{th}$  and  $j^{th}$  rows in the descriptor matrix and represent two compounds for which *S* is being computed;  $i, j = 1, 2, \dots, n$ , with  $n = 211$  for the PAH data set; and  $k = 1, 2, \dots, m$ , with  $m =$  number of descriptors, *d*. (The descriptor set used was height 0-3 signatures and all scalar descriptors, and values were range scaled so that they fell between 0 and 1.) From this, the dissimilarity matrix, *D*, was determined whose elements are complementary to the pairwise Tanimoto coefficients:

$$D(M_i, M_j) = 1 - S(M_i, M_j).$$

This dissimilarity matrix has been used with an iterative sphere-exclusion algorithm for calculating a single, deterministic test set,  $T_{te}$ , based on the assumption that a set of compounds spanning structure (descriptor) space will also span the property space<sup>50-52</sup>. A sphere-exclusion algorithm employs a user-defined threshold dissimilarity,  $t$ , to reject, in each iteration, all compounds that have a dissimilarity less than  $t$  with each compound in the test set for the current iteration. Thus,  $t$  is the radius of a hypersphere in the descriptor space and its value determines the size of  $T_{te}$ . In this study, this value was fixed at 0.03 because it translated to a reasonably sized  $T_{te}$  (23 compounds or about 10% of *n*).  $T_{te}$  was initialized by selecting the compound with the smallest sum of dissimilarities. Subsequent iterations saw the selection of the compound most similar to members of the test set for the current iteration.  $T_{te}$  was calculated in this way using D-SIM version 1.2<sup>53</sup>. The resulting training set,  $T_{tr}$ , comprising 188 compounds remaining after exclusion of the test set compounds from the original data set, was used in the subsequent robust QSPR modeling.

Interestingly, correlation analysis of  $\lambda$  with the sum of dissimilarities for the test and training set compounds seems to support the premise behind dissimilarity based compound selection via sphere-exclusion algorithms: structurally dissimilar compounds will also span the activity/property space. Thus, while the compounds of the  $T_{tr}$  training set exhibited no correlation with  $\lambda$  ( $r^2 = 0.0$ ), that with the 23 test set compounds was significant ( $r^2 = 0.5$ ).

## 4. Monte-Carlo cross-validation

It has been suggested that leave-one-out cross-validation is asymptotically inconsistent and tends to select an unnecessarily large model<sup>42</sup>. It has also been shown that Monte Carlo cross-validation is more stable and selects more optimal models than leave-one-out cross-validation<sup>42</sup>. In order to improve the PLS statistics, for the optimal combination of descriptors (height

0-3 signatures and  $\Delta\epsilon$ ), fractions of  $T_L$  have been randomly selected as training subsets. The number of molecules in each subset was varied from 95% of  $T_L$  to 5% in decrements of 5%. For each fraction: (a) 10,000 random subsets were generated resulting in a total of 190,000 models, and (b) the five best models (with highest  $q^2$  values) were identified resulting in a total of 95 models for final analysis.

- \* Electronic address: oavonli@sandia.gov
- <sup>1</sup> I. McCalloch, M. Heeney, C. Bailey, K. Genevicius, I. Macdonald, M. Shkunov, D. Sparrowe, S. Tierney, R. Wagner, W. Zhang, et al., *Nature Materials* 5, 328 (2006).
  - <sup>2</sup> J. Wu, W. Pisula, and K. Mullen, *Chem. Rev.* 107, 718 (2007).
  - <sup>3</sup> X. Feng, V. Marcon, W. Pisula, M. R. Hansen, J. Kirkpatrick, F. Grozema, D. Andrienko, K. Kremer, and K. Mullen, *Nature Materials* 8, 421 (2009).
  - <sup>4</sup> J. Nelson, J. Kwiatkowski, J. Kirkpatrick, and J. Frost, *Acc. Chem. Res.* 42, 1768 (2009).
  - <sup>5</sup> H. Yan, Z. Chen, Y. Zheng, C. Newman, J. Quinn, F. Dotz, M. Kastler, and A. Facchetti, *Nature* 457, 679 (2009).
  - <sup>6</sup> J. Bredas, J. Calbert, D. da Silva, and J. Cornil, *Proc. Nat. Acad. Sci.* 99, 5804 (2002).
  - <sup>7</sup> Y. Nagata and C. Lennartz, *J. Chem. Phys.* 129, 034709 (2008).
  - <sup>8</sup> J. J. Kwiatkowski, J. Nelson, H. Li, J. L. Bredas, W. Wenzel, and C. Lennartz, *Phys. Chem. Chem. Phys.* 10, 1852 (2008).
  - <sup>9</sup> D. L. Cheung and A. Troisi, *Phys. Chem. Chem. Phys.* 10, 5941 (2008).
  - <sup>10</sup> J. L. Bredas, D. Beljonne, V. Coropceanu, and J. Cornil, *Chem. Rev.* 104, 4971 (2004).
  - <sup>11</sup> R. Coehoorn, W. F. Pasveer, P. A. Bobbert, and M. A. J. Michels, *Phys. Rev. B* 72, 155206 (2005).
  - <sup>12</sup> V. Coropceanu, J. Cornil, D. A. da Silva, Y. Olivier, R. Silbey, and J. L. Bredas, *Chem. Rev.* 107, 2165 (2007).
  - <sup>13</sup> J. Kirkpatrick, V. Marcon, J. Nelson, K. Kremer, and D. Andrienko, *Phys. Rev. Lett.* 98, 227402 (2007).
  - <sup>14</sup> J. Kirkpatrick, V. Marcon, K. Kremer, J. Nelson, and D. Andrienko, *J. Chem. Phys.* 129, 094506 (2008).
  - <sup>15</sup> V. Marcon, J. Kirkpatrick, W. Pisula, and D. Andrienko, *Phys. Stat. Sci. (b)* 245, 820 (2008).
  - <sup>16</sup> V. Marcon, D. Breiby, W. Pisula, J. Dahl, J. Kirkpatrick, S. Patwardhan, F. Grozema, and D. Andrienko, *J. Am. Chem. Soc.* 131, 11426 (2009).
  - <sup>17</sup> Y. Olivier, L. Muccioli, V. Lemaire, Y. Geerts, C. Zannoni, and J. Cornil, *J. Phys. Chem. B* 113, 14102 (2009), URL <http://links.isiglobalnet2.com/gateway/Gateway.cgi?GWVersion=1&SrcAuth=KBib&SrcApp=KBib&KeyUI=000270911100009>.
  - <sup>18</sup> L. Schmidt-Mende, A. Fechtenkötter, K. Mullen, E. Moons, R. H. Friend, and J. D. MacKenzie, *Science* 293, 1119 (2001).
  - <sup>19</sup> J. Li, M. Kastler, W. Pisula, J. Robertson, D. Wasserfallen, A. Grimsdale, J. Wu, and K. Mullen, *Adv. Funct. Mat.* 17, 2528 (2007).
  - <sup>20</sup> R. A. Marcus, *Rev. Mod. Phys.* 65, 599 (1993).
  - <sup>21</sup> G. R. Hutchison, M. A. Ratner, and T. J. Marks, *J. Am. Chem. Soc.* 127, 2339 (2005).
  - <sup>22</sup> Y. Olivier, V. Lemaire, J. Bredas, and J. Cornil, *J. Phys. Chem. A* 110, 6356 (2006).
  - <sup>23</sup> W. Pisula, M. Kastler, D. Wasserfallen, T. Pakula, and K. Mullen, *J. Am. Chem. Soc.* 126, 8074 (2004).
  - <sup>24</sup> W. Pisula, Z. Tomovic, C. Simpson, M. Kastler, T. Pakula, and K. Mullen, *Chem. Mat.* 17, 4296 (2005).
  - <sup>25</sup> M. Kastler, W. Pisula, D. Wasserfallen, T. Pakula, and K. Mullen, *J. Am. Chem. Soc.* 127, 4286 (2005).
  - <sup>26</sup> P. J. Stevens, F. J. Devlin, C. F. Chabrowski, and M. J. Frisch, *J. Phys. Chem.* 98, 11623 (1993).
  - <sup>27</sup> M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, et al., *Computer code Gaussian 03, Revision B.05* (2003).
  - <sup>28</sup> J. Kirkpatrick, *Int. J. Quant. Chem.* 108, 51 (2008).
  - <sup>29</sup> J. L. Faulon, *J. Chem. Inf. Comp. Sci.* 34, 1204 (1994).
  - <sup>30</sup> J. Visco, R. S. Pophale, M. D. Rintoul, and J. L. Faulon, *J. Mol. Graph. Model.* 20, 429 (2002).
  - <sup>31</sup> <http://www.chemaxon.com>.
  - <sup>32</sup> D. Weininger, *Journal of Chemical Information and Computer Sciences* 28, 31 (1988), <http://pubs.acs.org/doi/pdf/10.1021/ci00057a005>, URL <http://pubs.acs.org/doi/abs/10.1021/ci00057a005>.
  - <sup>33</sup> P. Geerlings, F. D. Proft, and W. Langeneker, *Chem. Rev.* 103, 1793 (2003).
  - <sup>34</sup> O. A. von Lilienfeld and M. E. Tuckerman, *J. Chem. Phys.* 125, 154104 (2006).
  - <sup>35</sup> V. Marcon, O. A. von Lilienfeld, and D. Andrienko, *J. Chem. Phys.* 127, 064305 (2007).
  - <sup>36</sup> O. A. von Lilienfeld, *J. Chem. Phys.* 131, 164102 (2009).
  - <sup>37</sup> J. P. Perdew, R. G. Parr, M. Levy, and J. L. Balduz, *Phys. Rev. Lett.* 49, 1691 (1982).
  - <sup>38</sup> R. G. Parr and W. Yang, *Density functional theory of atoms and molecules* (Oxford Science Publications, 1989).
  - <sup>39</sup> P. Mori-Sánchez, A. J. Cohen, and W. Yang, *Phys. Rev. Lett.* 102, 066403 (2009).
  - <sup>40</sup> J. F. Janak, *Phys. Rev. B* 18, 7165 (1978).
  - <sup>41</sup> Y. Zhang and W. Yang, *J. Chem. Phys.* 109, 2604 (1998).
  - <sup>42</sup> J. Shao, *J. Acoust. Soc. America* 88, 486 (1993).
  - <sup>43</sup> A. Golbraikh and A. Tropsha, *J. Mol. Graph. Model.* 20, 269 (2002).
  - <sup>44</sup> MATLAB Copyright 1994-2008 by The MathWorks Inc.; 3 Apple Hill Drive, Natick, MA 07160-2098 USA.
  - <sup>45</sup> S. Wold, C. Albano, W. J. D. III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, and M. Sjostrom, in *Chemometrics: Mathematics and Statistics in Chemistry*, edited by B. R. Kowalski (D. Reidel Publishing Company,

- Dordrecht, Holland, 1984).
- <sup>46</sup> I. E. Frank and J. H. Friedman, *Technometrics* 35, 109 (1993).
- <sup>47</sup> <http://www.port.ac.uk/research/cmi/software/>.
- <sup>48</sup> D. C. Whitley, M. G. Ford, and D. J. Livingstone, *J. Chem. Inf. Comp. Sci.* 40, 1160 (2000).
- <sup>49</sup> P. M. Dean, *Molecular Similarity in Drug Design* (Springer, 2007).
- <sup>50</sup> A. M. Ferguson, D. E. Patterson, C. D. Garr, and T. L. Underiner, *J. Biomol. Screen.* 1, 65 (1996).
- <sup>51</sup> M. Snarey, N. K. Terrett, P. Willett, and D. J. Wilton, *J. Mol. Graph. Model.* 15, 372 (1997).
- <sup>52</sup> A. Golbraikh, M. Shen, Y. D. Xiao, K. H. Lee, and A. Tropsha, *J. Comput. Aided Mol. Des.* 17, 241 (2003).
- <sup>53</sup> M. Miara, M. M. Schveri, Q. Shi, X. Ye, E. Gruszecka-Kowalik, W. Bo, Z. Liu, H. M. Deutsch, and C. A. Venanzi (2010), in preparation.

## DISTRIBUTION

1	MS0899	Central Technical Files	9536 (electronic copy)
1	MS0123	D. Chavez, LDRD Office	1011 (electronic copy)



**Sandia National Laboratories**