Identifying Coarse-Grained Representations for Electronic Predictions

Chun-I Wang, J. Charlie Maier, and Nicholas E. Jackson*

Department of Chemistry, University of Illinois at Urbana-Champaign, 505 S Mathews Avenue, Urbana, Illinois, 61801, USA

E-mail: jacksonn@illinois.edu

Abstract

Coarse-grained (CG) simulations are an important computational tool in chemistry and materials science. Recently, systematic "bottom-up" CG models have been introduced to capture electronic structure variations of molecules and polymers at the CG resolution. However, the performance of these models is limited by the ability to select reduced representations that preserve electronic structure information, which remains a challenge. We propose two methods for (i) identifying important electronically coupled atomic degrees of freedom and (ii) scoring the efficacy of CG representations used in conjunction with CG electronic predictions. The first method is a physically-motivated approach that incorporates nuclear vibrations and electronic structure derived from simple quantum chemical calculations. We complement this physically-motivated approach with a machine learning technique based on the marginal contribution of nuclear degrees of freedom to electronic prediction accuracy using an equivariant graph neural network. By integrating these two approaches, we can both identify critical electronically coupled atomic coordinates and score the efficacy of arbitrary CG representations for making electronic predictions. We leverage this capability to make a connection

between optimized CG representations and the future potential for "bottom-up" development of simplified model Hamiltonians incorporating non-linear vibrational modes.

1 Introduction

Coarse-grained (CG) simulations have emerged as an essential computational tool in chemistry and materials science for simulating large systems that would be intractable at the atomistic resolution. In CG simulations, groups of atoms are combined into effective pseudoparticles via the definition of a CG representation (also called a CG "map"), reducing the number of nuclear degrees of freedom that must be simulated. Provided this map, several systematic methodologies have been introduced for establishing rigorous "bottom-up" correspondence between different simulation resolutions.[1–3] While these methodologies are well developed, the task of selecting the "good" CG maps remains challenging, with chemical intuition or convenience dictating most practical choices.[4–6] A number of systematic schemes for generating CG maps have been introduced that leverage essential dynamics,[7–10] graph theory,[11–14] and machine learning (ML),[15–18] but a dominant approach has yet to emerge.

In the context of modern physical chemistry, the ability to identify collective nuclear degrees of freedom that couple strongly to electronic subsystems of interest is of critical importance.[4,19] Collective configurational degrees of freedom motivates the widespread use of model Hamiltonians in the quantum dynamics community, with the modeling of excitation transport in photosynthetic complexes,[20–22] singlet fission in organic semiconductors,[23] charge and excitation transport along polymer chains,[24,25] and the absorption/emission spectra of molecular aggregates[26] utilizing a CG basis of sites with electron-phonon couplings in terms of linearized vibrational normal modes. In rare cases, couplings to non-linear degrees of freedom are known, but it is commonplace to select a single effective harmonic oscillator mode.[27,28] While several systematic frameworks exist for developing CG models from the bottom-up for classical simulations, no analogous techniques have been introduced for

CG model quantum mechanics (QM) Hamiltonians beyond the normal mode regime, which represents a notable gap in the field.

Recently, CG modeling methods have been introduced to facilitate the rigorous "bottom-up" coarse-graining of electronic structure variations in molecular systems, denoted Electronic Coarse-Graining (ECG).[19,29–34] If a CG map can be specified *a priori*, we have demonstrated the ability of heteroscedastic Gaussian Processes to reproduce the all-atom statistical distribution of electronic structure variations utilizing only the CG resolution.[33] While this systematic CG procedure is powerful, the task of identifying and selecting optimal CG maps for ECG is computationally laborious, requiring retraining of the ECG model for any new CG map.[33,34] Developing systematic and computationally efficient means of selecting CG maps are critically needed to advance ECG methods and ultimately connect with the "bottom-up" development of simplified QM model Hamiltonians for arbitrary molecular systems.

Here, we introduce two methods to identify high performing CG maps for use in electronic prediction models. First, we introduce a computationally efficient, physically-motivated scoring metric for CG maps that demonstrates strong correlations with the performance of ECG models. Second, we develop an approach to compute the marginal contributions of nuclear degrees of freedom to ECG model prediction accuracy, enabling the identification of critical nuclear degrees of freedom for inclusion in CG maps. We demonstrate the application of these techniques using a complex organic semiconductor molecule, 2-(4-methoxyphenyl)-7-octylbenzothienobenzothiophene (BTBT) employed in field-effect transistors that exhibits nontrivial electron-phonon couplings.[35] This work represents the first effort to develop systematic methods for identifying CG maps that capture electronically coupled collective nuclear degrees of freedom, forming the foundation for the systematic development of "bottom-up" CG procedures for simplified QM model Hamiltonians in future work.

2 Methods

2.1 Physically-Motivated Scoring Metric

We first introduce the physically-motivated approach for identifying nuclear degrees of freedom with strong electronic coupling, applicable at both the CG and atomistic resolutions. We define a score, D_i, associated with a particular CG bead, i (Eq. 1). D_i is defined as the modulus of a weighted sum (P_{jj} is a diagonal weight matrix) over atomic displacement vectors, $\vec{r}_j$, consistent with the CG map, M_{ij}, which is a linear operator that maps the atomic positions to the CG position of bead i utilizing linear weighting coefficients:

$$D_i = |M_{ij} P_{jj} \sum_{f=1}^{N_{\text{freq}}} \vec{r}_j^{\,f}|. \tag{1}$$

In Eq. 1, N_{freq} delineates the distinct frequencies associated with the frequency range over the averaging of displacement vectors occurs (see SI). A score, S, for a given CG map is obtained by summing over all individual CG bead scores, D_i. In Figure 1, a toy CG model of a hydrogen cyanide molecule illustrates the matrix expression of Eq. 1, in which the atoms are coarse-grained using two different maps.

The scoring metric, D_i, is inspired by the essential dynamics CG scheme[8,9] in that if the collective motions ($\vec{r}_j^{\,f}$) of multiple atoms contained within a given CG bead cancel out, this CG map will be a poor representation of the collective motion of the constituent configurational degrees of freedom. The larger the value of this summed displacement vector for a given CG bead, the more accurately collective atomistic motions are captured by the CG map. A simple example is provided in Figure 1b for a hypothetical linear molecule that demonstrates how the collective motions are preserved or cancelled out for two different CG representations. In this work, we assess the effectiveness of different forms of the CG mapping, M_{ij}, the weighting function, P_{jj}, and the displacement vector, $\vec{r}_j^{\,f}$ for capturing correlations between CG maps and electronic prediction capabilities. Hydrogen atoms are

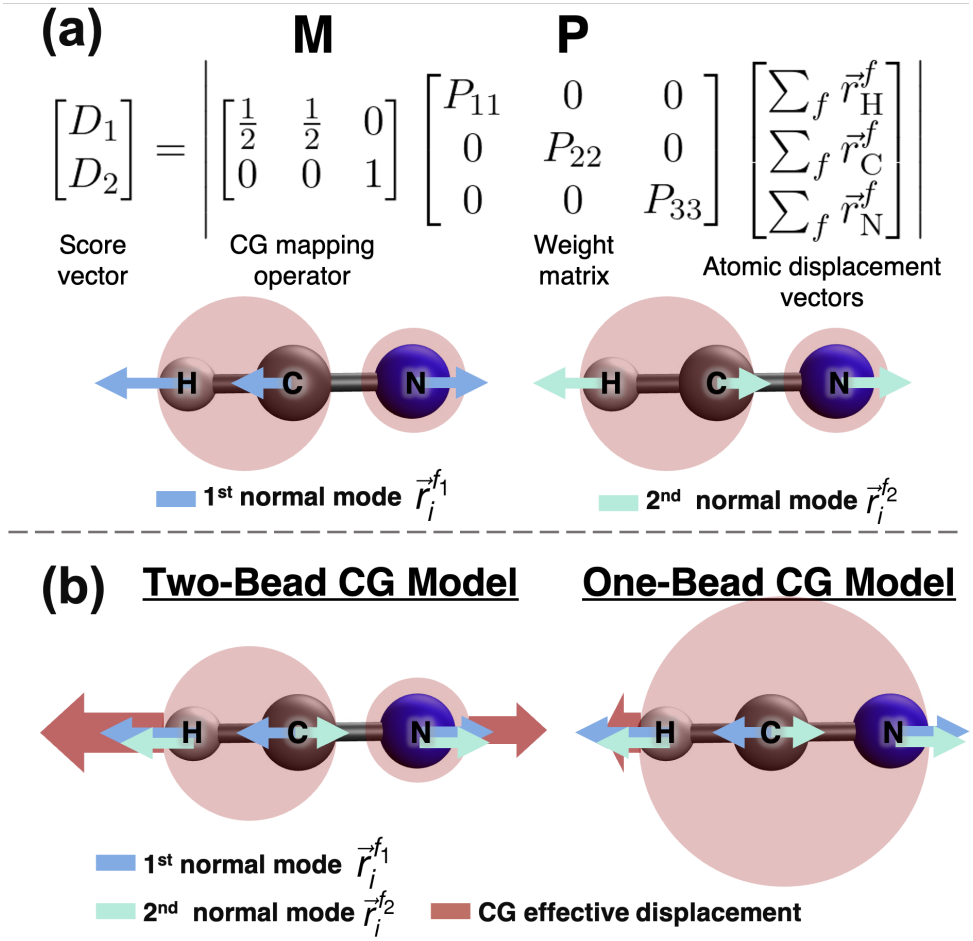


Figure 1: The physically-motivated approach for identifying and scoring CG maps using hydrogen cyanide (HCN) as a toy model along with two vibrational normal modes. (a) The matrix representation of Eq. 1 and the normal mode displacement of each atom. (b) Conceptual examples showing how collective motions are preserved in a two-bead CG model (left), and canceled out in a one-bead CG representation (right), the quantitative description of which is provided by Eq. 1.

excluded from all analysis. Note that if M_{ij} is the identity matrix then Eq. 1 acts as a scoring function at the atomistic resolution.

2.2 Data-Driven Evaluation Metrics

We complement this heuristic scoring method with a quantitative evaluation of the marginal contribution of atomic degrees of freedom to electronic prediction accuracy using equivariant graph neural networks (GNN).[36] The GNN is chosen as a data-driven metric for identifying the crucial atoms because it can incorporate both the molecular topology and the three-dimensional conformation during the learning process (see the SI for details). A training set of 12,000 molecular configurations were sampled via molecular dynamics (MD) simulations, with their electronic structure (highest occupied molecular orbital (HOMO) energy) characterized using density functional theory (DFT). Then, a separate instance of a GNN was trained for each heavy atom to be scored, with the coordinates of that particular atom removed from the input features to the GNN. Following training, the R^2 value of a testing dataset was compared to a reference GNN, in which all atoms were included in the input features. The score for any atom is calculated as the difference between the R^2 values for the reference GNN and the GNN with the atom excluded, and is interpreted as the marginal contribution of that atom to the relative variation in the HOMO energy of a molecular configuration. The results of this data-driven GNN model represent the most accurate characterization of the importance of atomic degrees of freedom to the target electronic prediction task.

To assess the efficacy of variations on the scoring metric (Eq. 1) for characterizing the electronic prediction capabilities of CG maps, we employ feedforward neural networks (FNN) to regress electronic properties (here the HOMO energy of BTBT) using only the CG representation.[29,37] 146,590 sampled MD configurations of BTBT are projected into a CG representation using a CG map (M_{ij}) and transformed into an inverse distance matrix, where the off-diagonal elements represent the reciprocal of the separation distance between CG parti-

cles. As the distance matrix is symmetric, we considered only the off-diagonal elements from the upper triangular portion as the FNN input feature—this featurization is the canonical featurization used in previous ECG models. An FNN model is then trained to predict the HOMO energy using only the CG representation of a molecule, and the R^2 ($S_{\mathrm{HOMO}}^{\mathrm{FNN-R2}}$) of this model serves as the ground truth for how accurately a given CG map captures electronic structure variations. We then correlate different scoring metrics ($S = \sum_i D_i$) derived using Eq. 1, with "good" CG map scoring metrics exhibiting strong linear correlation with $S_{\mathrm{HOMO}}^{\mathrm{FNN-R2}}$ (see the SI). All tested scoring metrics are dramatically cheaper than the cost required to train a FNN model to evaluate a CG map. Further details of model training are provided in the SI.

2.3 Dataset Construction

In this study, we employed MD simulations to sample a diverse set of BTBT monomer and dimer configurations at an atomistic resolution. The resulting two data sets comprise 146,590 monomer and 102,400 dimer structures, respectively, which were projected onto 107 different CG representations. Each CG configuration was then featurized using the inverse distance matrix associated with the respective CG representation and used as the input feature for FNN models. For GNN models, we randomly sampled 14,000 configurations from the monomer database (12,000 for training and 2,000 for testing) and used the atomic Cartesian coordinates as the input feature. The electronic properties targeted in this work were the HOMO energy of the BTBT monomer and the HOMO-HOMO intermolecular coupling of the BTBT dimer. These properties were calculated by density functional theory (DFT) calculations and the dimer projection technique[38–40] using the sampled atomistic configurations as input.

2.3.1 Collecting Molecular Configurations via MD Simulations

To obtain the BTBT configurations, we performed MD simulations of the nematic liquid crystal morphology at 555 K and 1.0 bar[35] using an OPLS-based force field.[41] We adapted the OPLS force field by reparameterizing the equilibrium geometry, atomic partial charges, and the torsional potentials of the inter-aromatic rings using DFT at the ωB97X-D3/cc-pVDZ level of theory. In Figure S2, we present a systematic characterization of the structural properties of BTBT in the four liquid-crystal phases (i.e., amorphous, nematic, smectic A, and smectic E phases), which exhibit consistent structural features and phase behavior with those observed in prior experimental and theoretical studies.[35,42–44] Further details on the parameterization and validation of the force field and the MD methodology can be found in the SI. With the obtained MD trajectory, we collected molecular pairs with center-of-mass distances ranging from 3 to 6 Å, resulting in a total of 102,400 pairs for the dataset of the HOMO-HOMO electronic coupling. From this set, we randomly selected 146,590 distinct monomer structures for the monomer dataset.

2.3.2 Characterization of Electronic Properties

The HOMO energy of each monomer structure sampled from MD was computed using DFT at the ωB97X-D3/cc-pVDZ level of theory. For molecular pairs, the electronic couplings were also computed at this level of theory utilizing the dimer projection approach.[38–40] In the dimer projection calculation, individual BTBT molecules were calculated in their neutral singlet states, and the off-diagonal Fock matrix elements were calculated as the coupling. In the present work, we aimed to investigate the hole transport property and thus calculated the off-diagonal Fock matrix element for the HOMO. The sign of the electronic coupling was determined using the phase-matching scheme introduced in previous work.[37,45] The distributions of the HOMO energy of the monomer dataset and the electronic coupling of the dimer dataset are shown in Figure S3. All DFT calculations were performed using ORCA[46] and we developed our own scripts to evaluate the electronic coupling. All computations for

the electronic characterization of the dataset were done using services provided by the OSG Consortium.[47–49]

3 Results and Discussion

3.1 Identifying Electronically Coupled Atomic Degrees of Freedom

First, we present the performance of the physically-motivated (Eq. 1) and GNN approaches for scoring the importance of atomic nuclei in BTBT for predicting its HOMO energy. Figure 2 shows the atomic scores derived via the marginal contributions from the GNN and three different forms of the scoring metric: MO (P_{ii} employing a normalized magnitude of the summed atomic HOMO density using Löwdin populations),[50] NM ($\vec{r}_j^f$ derived via normal mode analysis with DFT), and MO-NM (utilizing P_{ii} from the HOMO method, and $\vec{r}_j^f$ from the NM score). Figures 2b and 2c show that both the GNN and MO-NM approaches identify two carbon atoms located in the conjugated center of BTBT as the most important atomic coordinates influencing the HOMO energy. These carbon atoms have a high electron density associated with the HOMO, as revealed by the visualization of Löwdin population analysis in Figure 2d. Both approaches also highlight the contribution of carbon atoms located in the conjugated moiety adjacent to the methoxyphenyl group, which can be attributed to the rotational motion of the methoxyphenyl group that affects the electron delocalization via the overlap of adjoining carbon 2p orbitals. Importantly, these scoring metrics support the premise that the atomic importance provided by the quantitative GNN predictions can be qualitatively reproduced using physically-motivated metrics that exhibit lower computational cost than an ensemble of GNNs.

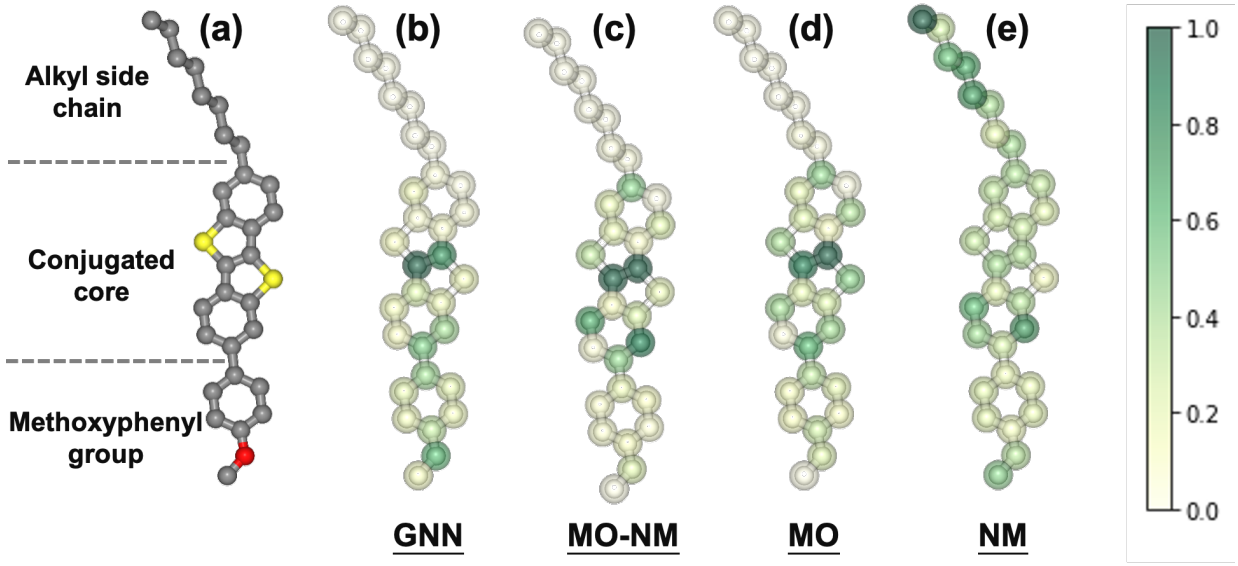


Figure 2: (a) Molecular structure of BTBT where hydrogen atoms are omitted for clarity. Normalized heatmaps showing the atomic scores evaluated by different methods: (b) graph neural networks, (c) MO-weighted normal mode displacement, (d) HOMO population, and (e) normal mode displacement. The atomic scores were normalized by the maximum atomic score of each evaluation metric. The heatmaps provide a visual representation of the relative importance of different atoms in BTBT, as evaluated by each metric.

3.2 Evaluation of CG Representations

The identification of electronically-coupled atomic coordinates provides a valuable guideline for developing CG representations that can accurately predict configurational variation of the HOMO energy for BTBT. Here, we evaluated a diverse set of CG representations for BTBT that were classified into four groups: two CG representations obtained from a systematic graph-based approach (denoted as GBCG),[14] 40 CG representations generated randomly (denoted as Random-1), 40 CG representations generated randomly that preserve the two carbon atoms located in the conjugated center of BTBT (denoted as Random-2), and 25 manually defined CG representations using the randomly generated CG models with higher metric scores as references (denoted as Improvement). Details of the random CG map generation algorithms are provided in the SI. Three evaluation metrics for all 107 generated CG maps are shown in Figure 3, with the score "S" denoting the sum over all D_i: (1) P_{ii} weighted by the marginal GNN scores and $\vec{r}_j^f$ by normal mode displacement vectors

($S_{\mathrm{HOMO}}^{\mathrm{GNN-NM}}$), (2) P_{ii} weighted by HOMO coefficients and $\vec{r}_j^{\,f}$ by normal mode displacement vectors ($S_{\mathrm{HOMO}}^{\mathrm{MO-NM}}$), and (3) the prediction accuracy of an FNN ECG model trained on the CG representation ($S_{\mathrm{HOMO}}^{\mathrm{FNN-R2}}$). Additional metrics that were explored are provided in the SI.

Across all three methods of evaluating CG representations, a general hierarchy of performance emerges amongst CG maps for the HOMO prediction task in BTBT: GBCG < Random-1 < Random-2 < Improvement. GBCG derived CG maps exhibit the lowest scores because no electronic information enters into the weighting functions and GBCG is biased toward homogeneous distribution for bead sizes when generating CG maps as illustrated in Figure 3d. Random-2 consistently outperformed Random-1 due to the pre-specified preservation of the two crucial carbon atoms in all randomly generated maps. The resolution of a CG map, which was quantified by the total number of CG particles (shown as a + mark in Figures 3a-c) had a considerable impact on all scores among the random mappings (Random-1 and Random-2), with higher resolution CG maps generally performing better than lower resolution ones as shown in Figure S5. Finally, CG maps hand-tailored (Improvement) utilizing the results of Figure 2 on average outperformed all other maps. For BTBT, in addition to retaining the two carbon atoms with the highest GNN or MO-NM scores, we found that atoms that are immediately pendant to the conjugated moieties should also be retained. These atoms include the carbon atom of the methoxyphenyl group connected to the conjugated core, the oxygen atom of the methoxyphenyl group, and the carbon atom linking the conjugated core and the alkyl side chain, which serves as a pivot point among the torsional motions. These results suggest that pre-existing knowledge of crucial atomic coordinates in electronic predictions can benefit the selection of optimal CG maps for electronic prediction tasks. Moreover, these results suggest that a heterogeneous distribution of CG bead resolutions may be more advantageous for electronic predictions than ones exhibiting a homogeneous distribution.

The analysis of CG scores also sheds light on a longstanding debate regarding the preservation of symmetry in CG modeling.[14,51–54] Among the 107 CG representations examined,

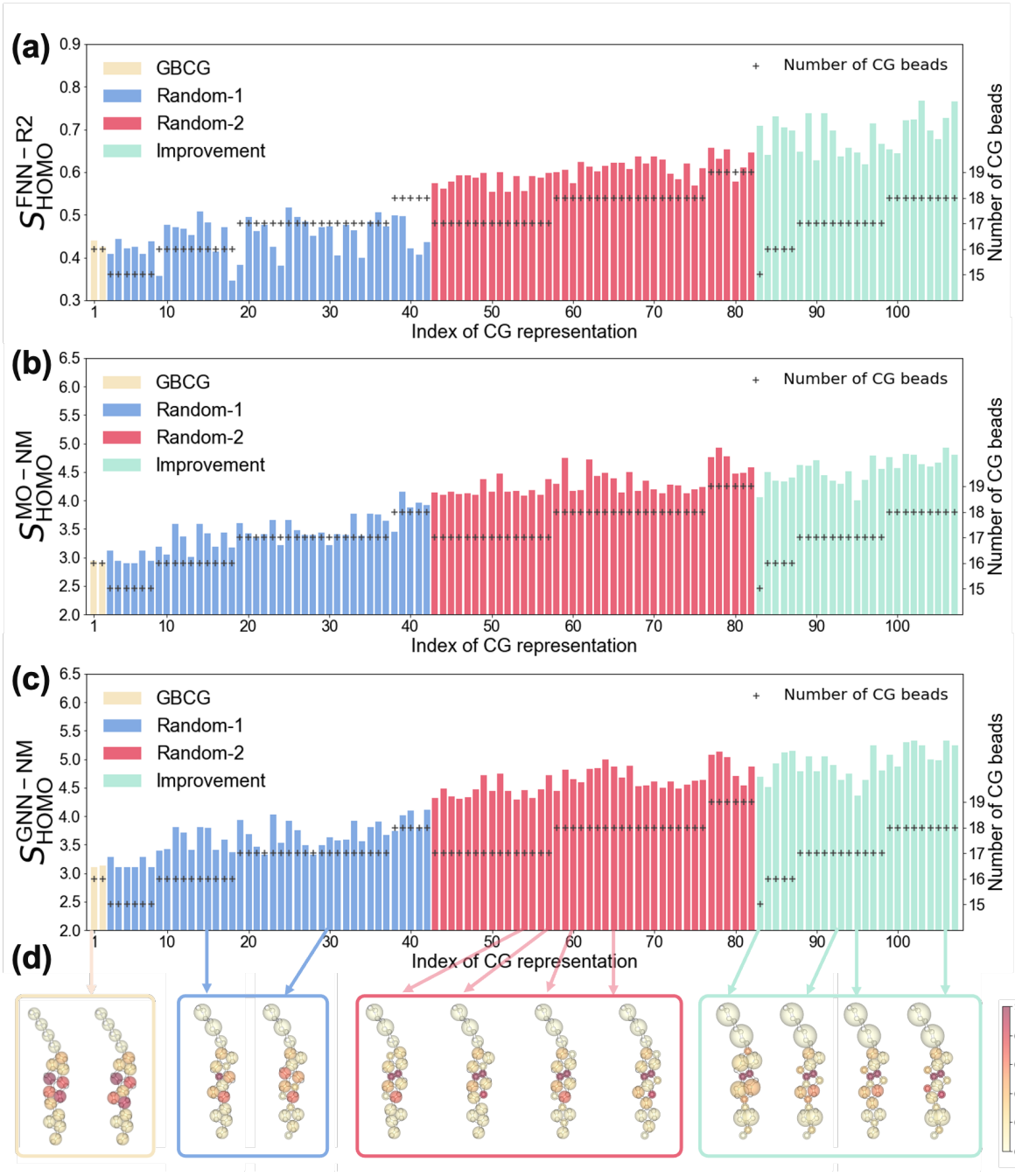


Figure 3: Global CG score of 107 CG representations, evaluated using two methods: (a) FNN model performance for HOMO energy prediction, (b) summation of the effective MO-NM displacement of all the CG particles, (c) summation of the normal mode displacement weighted by atomic scores of GNN metric shown in Figure 2b, and (d) selected CG representations in which the color of each particle indicates the normalized D_i score based on the MO-NM metric. The overall CG score provides a quantitative measure of the correspondence between each CG representation and HOMO energy. The three evaluation methods are compared to assess their ability to capture the relevant features of the system, as reflected in the CG score.

we observed a clear trend in the scores, particularly in relation to the degree of symmetry. Asymmetric CG mappings, such as the GBCG representations, consistently yielded lower scores compared to relatively symmetric ones, such as the top-10 of the Random-2 group (Figure S4). This suggests that "good representations," verified by both physically-motivated and data-driven metrics, are associated with a higher level of symmetry at the CG level. We believe that maintaining a symmetric CG topology is crucial for preserving the symmetry of CG bonded interactions, particularly in modeling conducting molecules that feature rigid or semi-flexible conjugated regions alongside dangling alkyl groups. Such symmetry not only helps mitigate numerical instability during iterative Boltzmann inversion but also ensures the accuracy and reliability of CG models.

While it is clear that information about atomic degrees of freedom and their relative importance to electronic prediction tasks can help inform the selection of CG maps, we now consider the computational cost and quantitative accuracies associated with CG maps selected via the different approaches. In Figures 4a-c, we show the correlations between $S_{\mathrm{HOMO}}^{\mathrm{MO-NM}}$, $S_{\mathrm{HOMO}}^{\mathrm{GNN-NM}}$, and $S_{\mathrm{HOMO}}^{\mathrm{NM}}$ and the ECG-derived prediction accuracy, denoted $S_{\mathrm{HOMO}}^{\mathrm{FNN-R2}}$. As a baseline, a scoring function utilizing only the normal mode displacement vectors and an identity matrix for the weighting matrix exhibits essentially no correlation with the predictive performance of ECG models (Figure 4a). This result is dramatically improved when utilizing the HOMO populations as the atomic weight matrix P_{ii}, resulting in a strong linear correlation with the performance of ECG models (Figure 4b). The scoring metric for CG maps is derived using the GNN marginal atomic contributions for P_{ii} in combination with normal modes for $\vec{r}_j^{\,f}$ (Figure 4c). A variety of alternative scoring metrics employing essential dynamics as estimates of $\vec{r}_j^{\,f}$, Huang-Rhys factor analysis for P_{ii}, and a variety of other approaches were also explored, with details provided in the SI. However, all such methods using Huang-Rhys factor analysis for P_{ii} exhibited significantly worse correlation with $S_{\mathrm{HOMO}}^{\mathrm{FNN-R2}}$, which represents the "ground-truth" of a CG map's predictive performance. The alternative metrics employing essential dynamics as the displacement vector ($\vec{r}_j^{\,f}$) show

moderate correlation with $S_{\mathrm{HOMO}}^{\mathrm{FNN-R2}}$ that indicates the robustness of Eq. 1 for quantifying the collective atomic motions within CG particles (see the SI for futher discussion). In fact, it can be observed that $S_{\mathrm{HOMO}}^{\mathrm{MO-NM}}$ correlates quite strongly with the "best" scoring metric which utilizes the GNN-derived marginal contribution weighting function (see Figure 4d). As obtaining $S_{\mathrm{HOMO}}^{\mathrm{GNN-NM}}$ requires training of an ensemble of GNN's over $\mathrm{O}(10^{4-5})$ electronic structure calculations, and $S_{\mathrm{HOMO}}^{\mathrm{MO-NM}}$ simply requires a normal mode quantum chemical (QC) analysis at the ground state geometry, the latter is recommended as a robust and practical methodology for selecting CG maps for ECG prediction models with the least computational cost.

With the establishment of the $S_{\mathrm{HOMO}}^{\mathrm{MO-NM}}$ score as a cost effective and quantitatively reliable means for evaluating the efficacy of CG maps for the HOMO prediction task in BTBT, we now examine the transferability of CG map scores to the off-target prediction property of the HOMO-HOMO intermolecular electronic coupling between BTBT dimers. We use an FNN to predict the intermolecular electronic coupling using only the CG resolution for the 107 CG maps previously examined. In Figure 5a, the R^2 scores of intermolecular electronic coupling for different CG presentations, $S_{\mathrm{Coupling}}^{\mathrm{FNN-R2}}$, were all above 0.9. Although there was a minor variation in $S_{\mathrm{Coupling}}^{\mathrm{FNN-R2}}$ for different CG representations, distinct separation between different classes of CG maps is still observed. Moreover, Figure 5c shows the correlation between $S_{\mathrm{Coupling}}^{\mathrm{FNN-R2}}$ and $S_{\mathrm{HOMO}}^{\mathrm{FNN-R2}}$, with a coefficient of determination of 0.607, which agrees with the physical intuition that CG maps optimized for MO energy prediction should work well for electronic couplings involving the same MO. These results indicate that the CG maps optimized for a single electronic property prediction task have the potential to be transferable between related prediction tasks.

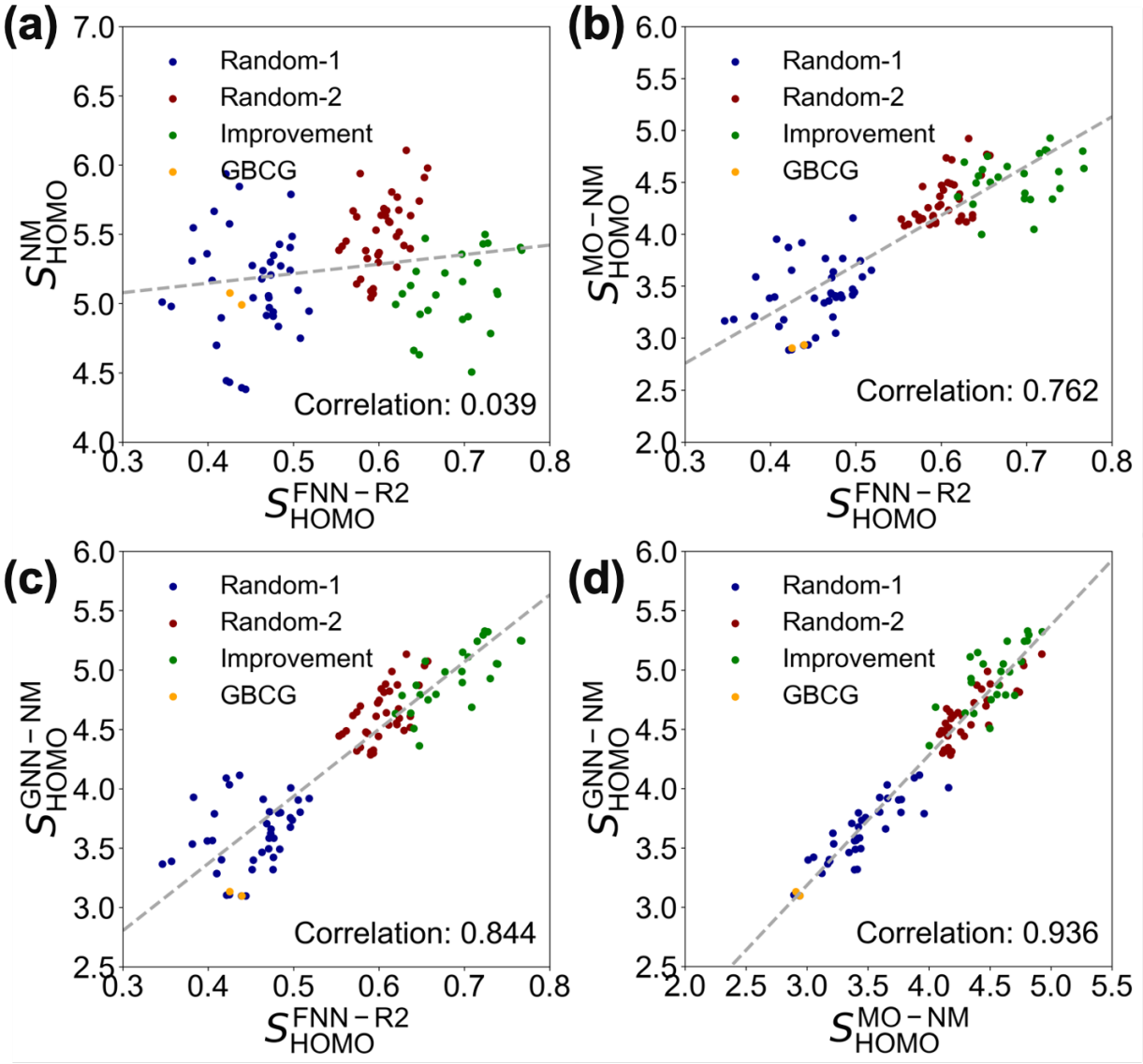


Figure 4: Correlation between different evaluation metrics for 107 CG representations: (a) sum over normal mode displacement without any weight ($S^{\mathrm{NM}}_{\mathrm{HOMO}}$) versus scores obtained from FNN ($S^{\mathrm{ML-R2}}_{\mathrm{HOMO}}$), (b) the normal mode scores weighted by HOMO population ($S^{\mathrm{MO-NM}}_{\mathrm{HOMO}}$) versus $S^{\mathrm{ML-R2}}_{\mathrm{HOMO}}$, (c) the normal mode scores weighted by GNN contribution ($S^{\mathrm{GNN-NM}}_{\mathrm{HOMO}}$) versus $S^{\mathrm{ML-R2}}_{\mathrm{HOMO}}$, and (d) versus $S^{\mathrm{MO-NM}}_{\mathrm{HOMO}}$ versus $S^{\mathrm{GNN-NM}}_{\mathrm{HOMO}}$

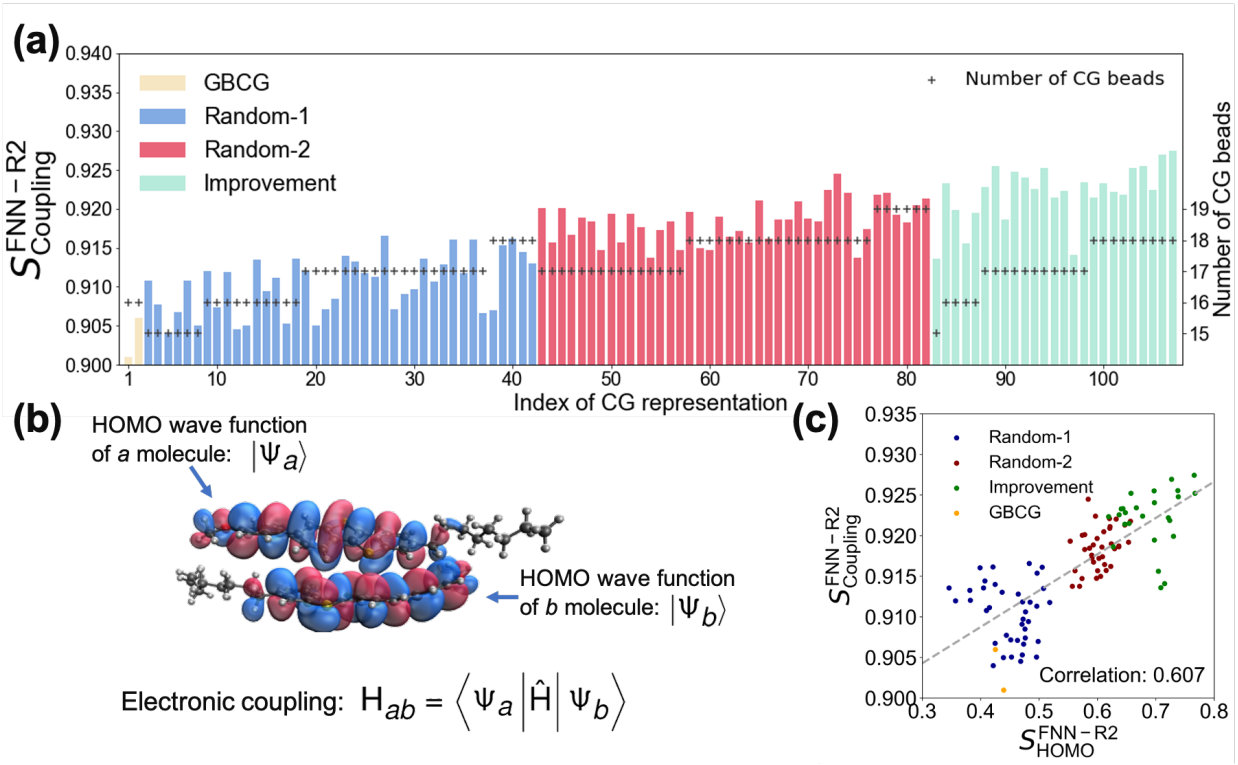


Figure 5: (a) Overall coarse-grained (CG) score of 107 CG representations evaluated by FNN model performance for electronic coupling prediction, (b) HOMO contour plots of a selected BTBT molecular pair, which illustrate the variation in HOMO density across the molecules and (c) scatter plot of CG scores based on FNN model performance for HOMO energy prediction and electronic coupling prediction for the 107 CG representations. The plots provide insight into the electronic properties of the molecules and can help guide the development of more accurate CG representations.

3.3 Transferability and Generality

The methods introduced herein should exhibit straightforward transferability to a variety of chemistries and electronic prediction properties of interest to the broader physical chemistry community. While initial tests with other molecular orbitals and their associated electronic properties shows a substantial robustness of the developed approach to any electronic target property, there is no expectation that the CG maps developed as a result of the methods introduced here should be transferable between arbitrary electronic prediction tasks. As the best CG maps derived in this study can be extracted using the HOMO population as a weighting function, there is no expectation that a CG map optimized for prediction tasks related to the HOMO should be accurate for prediction tasks related to other orbitals, particularly those with substantially different spatial structure of the orbitals. Such approaches have substantial promise in the organic semiconductor community but could also be of potential interest to the biological community interested in developing optimal CG maps that accurately capture e.g. hydrogen bonding or electronic polarization. As the scoring function defined by Eq. 1 is sufficiently general, users can experiment with any combination of weighting functions or displacement vectors relevant to their system of interest to access a computationally cheaper path to optimal CG makes than repeated training of a data-driven FNN in the traditional ECG approach. By eliminating the existing computational burden to the selection of "good" CG maps for electronic predictions to a standard ground state DFT normal mode analysis, this should significantly lower the barrier to the development of CG electronic prediction models in the broader community.

An important point of the developed methods relates to the utilization of the described approach for identifying electronically coupled vibrational modes, beyond the normal mode regime, of interest to quantum dynamics treatments. As the standard form of the Hamiltonian for quantum dynamical calculations employs bilinear couplings to, typically, localized normal mode vibrations, there is substantial interest moving forward in systematic approaches for identifying non-linear collective variables exhibiting strong couplings to elec-

tronic subsystems of interest. While it is common to identify electronically coupled normal modes by examining their associated Huang-Rhys factors, such approaches do not capture molecular motions beyond the normal mode regime. An interesting example of this occurs in the context of the BTBT molecule studied here, where the rotational motion of the phenyl dihedral angle is identified as being of critical importance to the HOMO energy (see Figure S3). Moreover, all scoring methods introduced here can immediately discriminate unimportant degrees of freedom, such as the alkylic side-chain in BTBT. By developing molecular representations utilizing only a subsystem of nuclear coordinates, there is the potential to use numerically driven techniques to identify their non-linear collective motions, which would be the first concrete step towards the establishment of electronically coupled non-linear vibrational modes. Moreover, the ability to achieve this for any arbitrarily complex chemistry from the "bottom-up" provides an interesting avenue to the systematic development of model Hamiltonians beyond the normal mode regime.

4 Conclusions

In this work, we presented two complementary methodologies for identifying strongly coupled atomic degrees of freedom and evaluating the effectiveness of CG maps utilized in electronic prediction tasks. A physically-motivated approach using normal mode analysis and quantum-chemically derived charge distributions was used to quantify electron-phonon couplings, demonstrating strong linear correlation with electronic predictions at CG resolutions generated using supervised machine learning. The proposed physically-motivated metric requires only a ground state normal mode analysis available in any quantum chemistry code, and is easy to implement. The implementation of this metric is available in our GitHub repository (https://github.com/TheJacksonLab/ECG_ScoringMetric). We have also introduced a data-driven approach leveraging equivariant GNNs to derive the marginal contribution of atomic degrees of freedom to electronic prediction performance, demonstrat-

ing good agreement with the physically-motivated approach. By leveraging both methods, we have demonstrated the potential to design CG models that preserve electronic-property-informative atomic coordinates for a targeted CG model resolution. Our results show the robustness and versatility of both methodologies in evaluating electronic properties at the CG level, suggesting that these approaches could lead to the development of optimized ECG models for a wide range of materials and systems, as well as "bottom-up" CG models for QM model Hamiltonians. We anticipate that these evaluation metrics will enable the systematic and automatic mapping of CG models that capture electronic structure or chemical reaction properties, facilitating the development of CG models that better reflect the behavior of complex molecular systems.

Acknowledgement

The development of the coarse-grained electronic models described in this article was supported by the National Science Foundation Chemical Theory, Models, and Computation division under award CHE-2154916. We acknowledge support from the Dreyfus Program for Machine Learning in the Chemical Sciences and Engineering during this project. The establishment of ML dataset was done using services provided by the OSG Consortium,[47–49] which is supported by the National Science Foundation awards #2030508 and #1836650.

Supporting Information Available

Computational detail of scoring metric D_i, Details of GNN and FNN scoring method, ML dataset establishment, creation of CG representations, and alternative scoring metrics for D_i.

References

(1) Lebold, K. M.; Noid, W. G. Dual approach for effective potentials that accurately model structure and energetics. *J. Chem. Phys.* **2019**, *150*, 234107.

(2) Pretti, E.; Shell, M. S. A microcanonical approach to temperature-transferable coarse-grained models using the relative entropy. *J. Chem. Phys.* **2021**, *155*, 094102.

(3) Dama, J. F.; Sinitskiy, A. V.; McCullagh, M.; Weare, J.; Roux, B.; Dinner, A. R.; Voth, G. A. The Theory of Ultra-Coarse-Graining. 1. General Principles. *J. Chem. Theory Comput.* **2013**, *9*, 2466–2480.

(4) Jackson, N. E. Coarse-Graining Organic Semiconductors: The Path to Multiscale Design. *J. Phys. Chem. B* **2021**, *125*, 485–496.

(5) Dhamankar, S.; Webb, M. A. Chemically specific coarse-graining of polymers: Methods and prospects. *J. Polym. Sci.* **2021**, *59*, 2613–2643.

(6) Jin, J.; Pak, A. J.; Durumeric, A. E. P.; Loose, T. D.; Voth, G. A. Bottom-up Coarse-Graining: Principles and Perspectives. *J. Chem. Theory Comput.* **2022**, *18*, 5759–5791.

(7) Arkhipov, A.; Freddolino, P. L.; Schulten, K. Stability and Dynamics of Virus Capsids Described by Coarse-Grained Modeling. *Structure* **2006**, *14*, 1767–1777.

(8) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules. *Biophys. J.* **2008**, *95*, 5073–5083.

(9) Zhang, Z.; Pfaendtner, J.; Grafmüller, A.; Voth, G. A. Defining Coarse-Grained Representations of Large Biomolecules and Biomolecular Complexes from Elastic Network Models. *Biophys. J.* **2009**, *97*, 2327–2337.

(10) Li, M.; Zhang, J. Z. H.; Xia, F. A new algorithm for construction of coarse-grained sites of large biomolecules. *J. Comput. Chem.* **2016**, *37*, 795–804.

(11) Gfeller, D.; De Los Rios, P. Spectral Coarse Graining of Complex Networks. *Phys. Rev. Lett.* **2007**, *99*, 038701.

(12) Gfeller, D.; De Los Rios, P. Spectral Coarse Graining and Synchronization in Oscillator Networks. *Phys. Rev. Lett.* **2008**, *100*, 174104.

(13) Chakraborty, M.; Xu, C.; White, A. D. Encoding and selecting coarse-grain mapping operators with hierarchical graphs. *J. Chem. Phys.* **2018**, *149*, 134106.

(14) Webb, M.; Delannoy, J.-Y.; de Pablo, J. J. Graph-Based Approach to Systematic Molecular Coarse-Graining. *J. Chem. Theory Comput.* **2019**, *15*, 1199–1208.

(15) Chen, Y.-L.; Habeck, M. Data-driven coarse graining of large biomolecular structures. *PLoS One* **2017**, *12*, e0183057.

(16) Wang, W.; Gómez-Bombarelli, R. Coarse-graining auto-encoders for molecular dynamics. *npj Comput. Mater.* **2019**, *5*, 1–9.

(17) Li, Z.; Wellawatte, G. P.; Chakraborty, M.; Gandhi, H. A.; Xu, C.; White, A. D. Graph neural network based coarse-grained mapping prediction. *Chem. Sci.* **2020**, *11*, 9524–9531.

(18) Ji, L.; Li, Y.; Wang, J.; Ning, A.; Zhang, N.; Liang, S.; He, J.; Zhang, T.; Qu, Z.; Gao, J. Community Reaction Network Reduction for Constructing a Coarse-Grained Representation of Combustion Reaction Mechanisms. *J. Chem. Inf. Model.* **2022**, *62*, 2352–2364.

(19) Wang, C.-I.; Jackson, N. E. Bringing Quantum Mechanics to Coarse-Grained Soft Materials Modeling. *Chem. Mater.* **2023**, *35*, 1470–1486.

(20) Runeson, J. E.; Lawrence, J. E.; Mannouch, J. R.; Richardson, J. O. Explaining the Efficiency of Photosynthesis: Quantum Uncertainty or Classical Vibrations? *J. Phys. Chem. Lett.* **2022**, *13*, 3392–3399.

(21) León-Montiel, R. d. J.; Kassal, I.; Torres, J. P. Importance of Excitation and Trapping Conditions in Photosynthetic Environment-Assisted Energy Transport. *J. Phys. Chem. B* **2014**, *118*, 10588–10594.

(22) Barclay, M. S.; Huff, J. S.; Pensack, R. D.; Davis, P. H.; Knowlton, W. B.; Yurke, B.; Dean, J. C.; Arpin, P. C.; Turner, D. B. Characterizing Mode Anharmonicity and Huang–Rhys Factors Using Models of Femtosecond Coherence Spectra. *J. Phys. Chem. Lett.* **2022**, *13*, 5413–5423.

(23) Chan, W.-L.; Berkelbach, T. C.; Provorse, M. R.; Monahan, N. R.; Tritsch, J. R.; Hybertsen, M. S.; Reichman, D. R.; Gao, J.; Zhu, X.-Y. The Quantum Coherent Mechanism for Singlet Fission: Experiment and Theory. *Acc. Chem. Res.* **2013**, *46*, 1321–1329.

(24) Dykstra, T. E.; Hennebicq, E.; Beljonne, D.; Gierschner, J.; Claudio, G.; Bittner, E. R.; Knoester, J.; Scholes, G. D. Conformational Disorder and Ultrafast Exciton Relaxation in PPV-family Conjugated Polymers. *J. Phys. Chem. B* **2009**, *113*, 656–667.

(25) Tozer, O. R.; Barford, W. Exciton Dynamics in Disordered Poly(p-phenylenevinylene). 1. Ultrafast Interconversion and Dynamical Localization. *J. Phys. Chem. A* **2012**, *116*, 10310–10318.

(26) Hestand, N. J.; Spano, F. C. Molecular Aggregate Photophysics beyond the Kasha Model: Novel Design Principles for Organic Materials. *Acc. Chem. Res.* **2017**, *50*, 341–350.

(27) Kun, H.; Avril, R. Theory of light absorption and non-radiative transitions in F-centres. *Proc. R. Soc. London A - Math. Phys. Sci.* **1950**, *204*, 406–423.

(28) Zhang, Y. Applications of Huang–Rhys theory in semiconductor optical spectroscopy. *J. Semicond.* **2019**, *40*, 091102.

(29) Jackson, N. E.; Bowen, A. S.; Antony, L. W.; Webb, M. A.; Vishwanath, V.; de Pablo, J. J. Electronic structure at coarse-grained resolutions from supervised machine learning. *Sci. Adv.* **2019**, *5*, eaav1190.

(30) Jackson, N. E.; Bowen, A. S.; de Pablo, J. J. Efficient Multiscale Optoelectronic Prediction for Conjugated Polymers. *Macromolecules* **2020**, *53*, 482–490.

(31) Simine, L.; Allen, T. C.; Rossky, P. J. Predicting optical spectra for optoelectronic polymers using coarse-grained models and recurrent neural networks. *Proc. Natl. Acad. Sci.* **2020**, *117*, 13945–13948.

(32) Sivaraman, G.; Jackson, N. E. Coarse-Grained Density Functional Theory Predictions via Deep Kernel Learning. *J. Chem. Theory Comput.* **2022**, *18*, 1129–1141.

(33) Maier, J. C.; Jackson, N. E. Bypassing backmapping: Coarse-grained electronic property distributions using heteroscedastic Gaussian processes. *J. Chem. Phys.* **2022**, *157*, 174102.

(34) Alessandri, R.; de Pablo, J. J. Prediction of Electronic Properties of Radical-Containing Polymers at Coarse-Grained Resolutions. *arXiv* **2022**, 2209.02072 [cond–mat.soft].

(35) Han, M. J.; Wei, D.; Kim, Y. H.; Ahn, H.; Shin, T. J.; Clark, N. A.; Walba, D. M.; Yoon, D. K. Highly Oriented Liquid Crystal Semiconductor for Organic Field-Effect Transistors. *ACS Cent. Sci.* **2018**, *4*, 1495–1502.

(36) Wang, L.; Liu, Y.; Lin, Y.; Liu, H.; Ji, S. ComENet: Towards Complete and Efficient Message Passing for 3D Molecular Graphs. 2022.

(37) Wang, C.-I.; Joanito, I.; Lan, C.-F.; Hsu, C.-P. Artificial neural networks for predicting charge transfer coupling. *J. Chem. Phys.* **2020**, *153*, 214113.

(38) Ohta, K.; Closs, G. L.; Morokuma, K.; Green, N. J. Stereoelectronic effects in intramolecular long-distance electron transfer in radical anions as predicted by ab-initio MO calculations. *J. Am. Chem. Soc.* **1986**, *108*, 1319–1320.

(39) Senthilkumar, K.; Grozema, F. C.; Bickelhaupt, F. M.; Siebbeles, L. D. A. Charge transport in columnar stacked triphenylenes: Effects of conformational fluctuations on charge transfer integrals and site energies. *J. Chem. Phys.* **2003**, *119*, 9809–9817.

(40) Valeev, E. F.; Coropceanu, V.; da Silva Filho, D. A.; Salman, S.; Brédas, J.-L. Effect of Electronic Polarization on Charge-Transport Parameters in Molecular Organic Semiconductors. *J. Am. Chem. Soc.* **2006**, *128*, 9882–9886.

(41) William L. Jorgensen, a.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *1996*, ,45.

(42) Inoue, S.; Minemawari, H.; Tsutsumi, J.; Chikamatsu, M.; Yamada, T.; Horiuchi, S.; Tanaka, M.; Kumai, R.; Yoneya, M.; Hasegawa, T. Effects of Substituted Alkyl Chain Length on Solution-Processable Layered Organic Semiconductor Crystals. *Chem. Mater.* **2015**, *27*, 3809–3812.

(43) Yoneya, M.; Minemawari, H.; Yamada, T.; Hasegawa, T. Interface-Mediated Self-Assembly in Inkjet Printing of Single-Crystal Organic Semiconductor Films. *J. Phys. Chem. C* **2017**, *121*, 8796–8803.

(44) Okamoto, T.; Yu, C. P.; Mitsui, C.; Yamagishi, M.; Ishii, H.; Takeya, J. Bent-Shaped p-Type Small-Molecule Organic Semiconductors: A Molecular Design Strategy for Next-Generation Practical Applications. *J. Am. Chem. Soc.* **2020**, *142*, 9083–9096.

(45) Wang, C.-I.; Braza, M. K. E.; Claudio, G. C.; Nellas, R. B.; Hsu, C.-P. Machine Learning for Predicting Electron Transfer Coupling. *J. Phys. Chem. A* **2019**, *123*, 7792–7802.

(46) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *J. Chem. Phys.* **2020**, *152*.

(47) Pordes, R.; Petravick, D.; Kramer, B.; Olson, D.; Livny, M.; Roy, A.; Avery, P.; Blackburn, K.; Wenaus, T.; Würthwein, F.; Foster, I.; Gardner, R.; Wilde, M.; Blatecky, A.; McGee, J.; Quick, R. The open science grid. J. Phys. Conf. Ser. 2007; p 012057.

(48) Sfiligoi, I.; Bradley, D. C.; Holzman, B.; Mhashilkar, P.; Padhi, S.; Wurthwein, F. The pilot way to grid resources using glideinWMS. 2009 WRI World Congress on Computer Science and Information Engineering. 2009; pp 428–432.

(49) OSG, OSPool. 2006; `https://osg-htc.org/services/open_science_pool.html`.

(50) Szabó, A.; Ostlund, N. S. *Modern quantum chemistry : introduction to advanced electronic structure theory*; Dover publications: Mineola, NY, USA, 1996; pp 138–152.

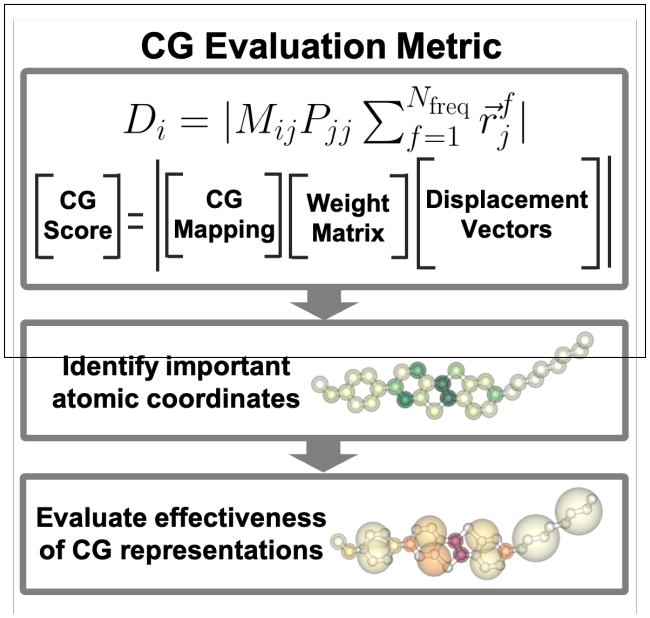
(51) Cao, Z.; Voth, G. A. The multiscale coarse-graining method. XI. Accurate interactions based on the centers of charge of coarse-grained sites. *J. Chem. Phys.* **2015**, *143*.

(52) Khot, A.; Shiring, S. B.; Savoie, B. M. Evidence of information limitations in coarse-grained models. *J. Chem. Phys.* **2019**, *151*.

(53) Durumeric, A. E. P.; Voth, G. A. Adversarial-residual-coarse-graining: Applying machine learning theory to systematic molecular coarse-graining. *J. Chem. Phys.* **2019**, *151*.

(54) Chakraborty, M.; Xu, J.; White, A. D. Is preservation of symmetry necessary for coarse-graining? *Phys. Chem. Chem. Phys.* **2020**, *22*, 14998–15005.

TOC Graphic


CG Evaluation Metric
$D_i = |M_{ij} P_{jj} \sum_{f=1}^{N_{\text{freq}}} \vec{r}_j^{\,f}|$
CG Score
=
CG Mapping
Weight Matrix
Displacement Vectors
Identify important atomic coordinates
Evaluate effectiveness of CG representations