OCNet: A Domain Knowledge-Enhanced General Moleculer Representation Framework for Optoelectronic and Charge-transport Materials

Guojiang Zhao,[†,‖] Qi Ou,[‡,‖] Zifeng Zhao,[*,†,‖] Shangqian Chen,[†] Haitao Lin,[†] Xiaohong Ji,[†] Zhen Wang,[†] Hongshuai Wang,[†] Hengxing Cai,[†] Lirong Wu,[†] Shuqi Lu,[†] FengTianCi Yang,[†] Zhifeng Gao,[*,†] and Zheng Cheng[*,¶,§]

†*DP Technology, Beijing 100080, P.R. China*

‡*SINOPEC Research Institute of Petroleum Processing Co., Ltd, Beijing 100083, China.*

¶*School of Mathematical Sciences, Peking University, Beijing 100871 China*

§*AI for Science Institute, Beijing 100084, P.R. China*

‖*Contributed equally to this work*

E-mail: zhaozf@aisi.ac.cn; gaozf@dp.tech; chengz@aisi.ac.cn

Abstract

The characterization of material properties plays a crucial role in revealing the structure-property relationship and optimizing device performance. Organic optoelectronic and transporting materials, widely used in various fields, face challenges in experimental property characterization not only due to their high cost but also the requirement of multidisciplinary knowledge. To address this problem, we introduce OCNet, a domain knowledge-enhanced representation learning framework, with which the efficient and accurate virtual characterization is made possible. Based on the SE(3) transformer architecture and a self-constructed large-scale conjugated molecular

database with millions of structures and properties, OCNet realizes general molecular and bimolecular representation and supports the integration of domain knowledge features. In multiple optoelectronic property prediction tasks, OCNet shows a significant improvement in accuracy compared to previously reported models. It also constructs a DFT-accuracy database for the transfer integrals of thin-film materials and renders the general prediction of such properties possible. With its user-friendly interface, OCNet can serve as an effective virtual characterization tool, facilitating the development of optoelectronic devices and other functional material research.

I. Introduction

The characterization of material properties serves as a pivotal link in deciphering the structure-property relationship and holds the key to optimizing the device performance. For functional materials, specific property requirements vary across different application scenarios, making comprehensive property characterization of utmost importance.[1] However, such processes demand multidisciplinary knowledge, professional operational expertise, costly instrumentation, and significant time investment, which have posed a substantial obstacle to the development of functional materials.

Organic optoelectronic materials, a burgeoning frontier in functional materials, have found extensive applications in diverse academic and commercial arenas such as display,[2] clean energy,[3] and biosensing,[4] owing to their unique tunability in optoelectronic and charge transport characteristics. These materials typically comprise conjugated molecules and are utilized in the form of solid films as the core constituents of optoelectronic devices. Experimentally characterizing the optoelectronic and charge transport properties of molecules in solution and/or solid film and elucidating their influence on the performance of various devices is crucial for driving the advancement of the organic optoelectronics industry.[5] Nevertheless, as previously noted, the associated property characterization costs and the prerequisite for researchers' multidisciplinary knowledge and technical proficiency have become

hurdles in the development of novel devices. Hence, an open question emerges: can we devise virtual characterization tools to acquire relatively accurate material properties at a reduced cost, thereby substantially diminishing the reliance on experimental characterization during device development?

Attempts from various levels of theory have been made to address this question. On the one hand, at the molecular level, traditional quantum chemistry methods can delineate the electronic relaxation processes among different electronic states based on the geometric configuration of a single molecule, enabling relatively accurate predictions on the material's optoelectronic properties.[6–11] The feasibility of such method actually lies in the fact that the intermolecular interaction between organic molecules is weak.[12] By leveraging the geometric configuration of bimolecules in films, the electronic coupling (transfer integral) between molecules can be described to assess the electron transport probability.[13,14] However, the aforementioned simulation methods entail high computational demands as well as master specialized computational techniques of the researcher, such as the selection of functionals and basis sets and the preparation of force field topology files. On the other hand, data-driven methods have exhibited significant potential in predicting computational or experimentally characterized properties with high accuracy and efficiency.[15–22] Yet, these methods usually rely on feature engineering predicated on expert domain knowledge, often suffering from poor transferability to dissimilar property tasks and being at a disadvantage when dealing with high-diversity databases.

Recent deep learning approaches for optoelectronic property prediction primarily extract a series of predefined atom and bond features, such as the formal charge on the atom, from the 2D graph of the molecule and construct the molecular representation via the message passing mechanism.[23–26] Currently, the state-of-the-art (SOTA) accuracy has been attained in multiple optoelectronic property prediction tasks.[24–26] However, this 2D graph convolution method disregards 3D information and restricts its applicability to tasks involving 3D information, such as predicting the transfer integral(TI) of the film system based on the geometry

of bimolecules. In the prediction of charge transport properties, Bhat et al. constructed a transfer integral database using a crystal structure library and predicted the transfer integral in a crystal environment based on the 3D graph convolution molecular representation.[13] Owing to the dearth of experimental data and certain force field parameters in the molecular dynamic (MD) simulations of conjugated systems, no structural database of optoelectronic films has been reported thus far. For film systems, most existing studies predict the transfer integral of a single system based on the Coulomb matrix.[27–29] In principle, deep learning can yield a molecular representation comparable to that obtained through domain expert feature engineering by pre-training on relevant large-scale (usually over 3M) 3D positions or low-precision labels within the field. This paradigm has been demonstrated to outperform previous methods in data-rich scenarios such as large language models,[30–32] small molecule drug property prediction,[33–35] and general force fields of inorganic materials.[36–38]

To achieve accurate and efficient virtual characterization of functional materials considering both optoelectronic and transport properties, we introduce OCNet to bridge the gap in achieving general and low data-requirement molecular representations of these properties. Firstly, we construct a large-scale conjugated molecular database containing 12M GFN2-xTB[39] optimized structures and sTDA-xTB[40] level optoelectronic properties, together with a conjugated bimolecular database with 9M bimolecular conformations collected from 100K molecular dynamic(MD) simulations of organic films and transfer integrals with GFN1-xTB[41] accuracy. Based on these two large-scale databases, our OCNet realizes a general conjugated molecular representation and a bimolecular representation using the SE(3) Transformer architecture.[33] Additionally, for specific downstream property prediction tasks, we support domain experts in fusing domain knowledge features with molecular representations to enhance the interpretability of OCNet and further improve its performance on scarce datasets. The accuracy of our OCNet on various computationally and experimentally labeled properties is remarkably enhanced compared to the previously reported models, attesting to the superiority of our general molecular representation. Moreover, we also establish

a bimolecular transfer integral dataset containing 1.8M data points computed via density functional theory (DFT) based on 45K molecular films. The accuracy performance of our OCNet on this dataset can be exploited for subsequent quantitative prediction of film transport properties, marking a significant advancement in film transfer integral prediction from single-system modeling to general model prediction. Finally, to make the OCNet more accessible to the broader research community, we provide both the OCNet code, pre-training database and corresponding molecular and bimolecular representation online. We also provide a user-friendly web application to predict downstream properties in this work. With the free and user-friendly accessibility of OCNet, we anticipate that the general molecular representation of OCNet will serve as an effective tool for the virtual characterization of the optoelectronic and transport properties of organic optoelectronic materials, expediting the development of optoelectronic devices and demonstrating the potential for application in other functional material research scenarios related to organic conjugated molecules.

II. Results and Discussion

Overview of OCNet framework

Our OCNet framework (Figure 1) primarily consists of two modules: one is the general and low-data-requirement conjugated molecular and bimolecular representations, and the other is for downstream task property prediction, which integrates domain features with general molecular representations.

General molecular and bimolecular representations for conjugated systems

Due to the lack of large-scale databases for pre-training conjugated molecular and bimolecular representations, we construct a conjugated molecular and bimolecular database that contains 10 million(M) level geometries and corresponding optoelectronic properties or transfer integrals at the tight-binding (TB) level. Our database includes 16 different elements (H, B, C,

N, O, F, Si, P, S, Cl, Br, I, Ir, Ge, Se, As) and three distinct types of conjugated molecules (metal-organic complexes, condensed hetero-polycyclic aromatic molecules, and fragment-assembly-based aromatic molecules), thereby covering a broad range of the chemical space of organic conjugated systems. We illustrate the construction process of the molecular and bimolecular database in Figure 1a.

Firstly, we collect the Ir complex open-source dataset[42] containing 0.84M structures and the COMPAS-2x open-source dataset[43] comprising 0.5 M cata-condensed hetero-polycyclic aromatic molecules. Then, we generate an additional 11M molecular structures using ring fusion and fragment assembly methods (detailed in the Methods section, Figures S1 and S2). When generating million-level condensed poly(hetero)cyclic aromatic molecules using the ring fusion method, we allow carbon or heteroatoms to be shared by two or three rings (Figure S1b). This approach makes our database the first to include molecules with multiple resonance structures. Moreover, we also generate millions of fragment-assembly-based molecules by connecting carbon atoms across different conjugated fragments for the first time. To further demonstrate the diversity of our molecular database, we compare our molecular database with the open-source COMPAS-2x dataset in terms of heavy atom numbers (Figure 2a, b) and molecular weight (Figure S3) distribution. It can be observed that our molecular database contains larger molecules: while most molecules in COMPAS-2x have fewer than 50 heavy atoms and a molecular weight of less than 600 Da, approximately 60% of the molecules in our database have more than 50 heavy atoms and a molecular weight greater than 600. We also compare the distribution of our molecular database with the COMPAS-2x database and the largest fragment-assembly-based database, FORMED[44] with 0.1M structures, in the 2D space of our molecular representation using t-SNE analysis(Figure 2c). It can be observed that COMPAS2 and FORMED only occupy partial regions of our molecular database, which further indicates that our molecular data is widely distributed across the chemical space of conjugated systems.

We also sample 9.5 dimer conformations and corresponding TB-level HOMO-HOMO and

LUMO-LUMO transfer integrals from 100 K molecular films as our bimolecular database. To the best of our knowledge, this is the first bimolecular dataset under film conditions. To achieve this, we first select molecules with relatively low electron or hole reorganization energies at the GFN2-xTB level from our molecular database. Then, we obtain the film structures through MD simulations using the force field we develop for organic conjugated molecules (detailed in the methodology section). We also demonstrate the distribution of heavy atoms and molecular weight in our bimolecular database (Figure S4). It can be seen that the maximum heavy atom count reaches 350, and the maximum molecular weight reaches 9000 Da, indicating that our bimolecular database possesses significant molecular diversity.

To ensure generality, we only use the system's elements and distance kernel matrix as the initial atomic and pair representations. These representations are then aggregated into a global molecular representation through 15 encoder layers using the attention mechanism (detailed in Figure S5 and the Methods section). In this work, we choose a SE(3) Transformer architecture which contains 15 encoder layers to construct molecular representation learning (MRL) models, as it has demonstrated high performance on small organic and drug molecules.[33,45]Then, we optimize all the weight parameters of the molecular representation learning (MRL) models through pre-training on our molecular and bimolecular databases(Figure 1b). By learning from large amounts of unlabeled structures and TB-level optoelectronic properties and transfer integral data, we can significantly reduce the data requirements of our molecular and bimolecular representation when fine-tuning downstream optoelectronic and transport-related tasks.

Domain feature integrated molecular representations for downstream property prediction

For specific downstream tasks, OCNet provides two approaches for virtual characterization: directly modeling the downstream properties based on molecular representations or inte-

grating prior knowledge to further enhance the molecular representation's capability and interpretability. Specifically, we use multilayer perceptrons to extract both molecular representations and domain-specific features, which are then fused into a single representation for downstream property prediction (Figure 1c, and detailed in the methodology section). To comprehensively demonstrate the performance of OCNet in different downstream scenarios, we evaluate its accuracy on QM-calculated or experiment-measured optoelectronic properties and transfer integrals in crystal or film environments. For all property prediction tasks, we set the training-to-test ratio at 8:2 and used mean absolute error (MAE) and the coefficient of determination (R^2) to evaluate model accuracy. Moreover, to make the OCNet more accessible to the broader research community, we also integrated these downstream property prediction models into a user-friendly interactive web app(https://funmg.dp.tech/) for material design in various scenarios(Figure 1d), such as photovoltaic and display.

Performance of OCNet on optoelectronic task

We utilize the open-source OCELOT chromophore dataset,[26,46] which provides DFT and TDDFT-computed data, to evaluate the accuracy of OCNet in predicting QM-calculated optoelectronic properties. Specifically, we compare the performance of OCNet with reported SOTA models in terms of the HOMO-LUMO gap (H-L), lowest-lying singlet excitation energies (S0-S1), electron reorganization energies (ER), and hole reorganization energies (HR) (Figure 3a). We define the accuracy score of OCNet on the dataset as the ratio of the MAE of the SOTA model to the MAE of OCNet. It can be observed that OCNet achieves the highest accuracy across all four property prediction tasks, with accuracy scores improving by at least 15% for all tasks and by 60% for the HR prediction task. We also demonstrate the correlation between the OCNet predictions and the corresponding QM-calculated values for all four properties(Figure 3d). OCNet successfully achieves quantitative predictions for the S0-S1 and H-L tasks, with MAEs of 0.199 eV and 0.008 eV, and R^2 values of 0.803 and 0.987, respectively. OCNet achieves MAEs of 0.082 eV and 0.087 eV for the ER and HR prediction

tasks, with R^2 values of 0.575 and 0.511, respectively. This level of accuracy is sufficient for screening molecules with low reorganization energies. We further compare the MAE and R^2 of the reported SOTA model, as well as OCNet with(w/) and without(w/o) pre-training, across all four properties (Table S1 and S2). The results show that pre-training on large-scale data significantly enhances the accuracy of 3D representations. Due to the integration of prior knowledge, such as atom types and whether chemical bonds are π-conjugated, 2D graph convolutional networks achieve higher accuracy than OCNet w/o pre-training for all four property prediction tasks. However, they are consistently underperformed by OCNet w/ pre-training. For instance, when predicting the S0-S1 property, the reported SOTA model achieves a MAE of 0.249 eV and a R^2 of 0.760, while OCNet w/o pre-training has a MAE of 0.318 eV and a R^2 of 0.544. In contrast, OCNet w/ pre-training achieves a significantly improved MAE of 0.199 eV and a R^2 of 0.803. All these results demonstrate that only by adopting a large-scale data pre-training strategy can 3D molecular representations outperform previous 2D GNN models.

Subsequently, we evaluate the accuracy of OCNet on the experimental optoelectronic dataset, Deep4Chem.[47] Given the complexity of the solution environment, reported SOTA models typically employ strategies such as directed message passing, introducing additional edge and subgraph information or integrating domain prior knowledge to enhance molecular representation.[24,25] We integrate domain-specific features utilized in SuboptGraph[25] into our molecular representation to predict absorption wavelength (Abs.), emission wavelength (Emi.), photoluminescence quantum yield (PLQY), and full width at half maximum (FWHM). We compare the performance of OCNet with SOTA models in predicting these four properties (Figure 3 b). The accuracy score of OCNet on the datasets is defined as the ratio of the MAE of the SOTA model to the MAE of OCNet. Although SOTA 2D graph convolutional models incorporate additional edge and subgraph information to enhance molecular representation, our 3D geometry-based OCNet achieves improvements of 18% and 13% in accuracy score for the prediction of absorption and emission wavelengths,

respectively. Since our pre-training does not include tasks related to PLQY or FWHM, the performance improvement of OCNet for these properties is relatively modest, approximately 5%. We also demonstrate the correlation between the OCNet predicted values and experimental values(Figure 3e). OCNet accurately predicts absorption and emission wavelengths, achieving MAEs of 7.085 nm and 11.167 nm, respectively, with corresponding R^2 values of 0.982 and 0.949. The accuracy of OCNet in predicting PLQY and FWHM is also capable of screening molecules with high PLQY and narrow emission, with MAEs of 0.101 and 9.123 nm, and R^2 values of 0.722 and 0.719. We further compare the MAEs and R^2 values of the reported models (Uni-Mol and reported SOTA model SuboptGraph), along with OCNet w/ and w/o pre-training, for absorption and emission wavelength predictions (Figure 3c, Tables S3 and S4). Due to significant structural differences between drug molecules and conjugated molecules, Uni-Mol, which is pre-trained on drug molecule conformations, exhibits relatively poor accuracy in absorption and emission wavelength prediction tasks. Specifically, the MAE for emission energy prediction reaches 16 nm, whereas the MAE for the other models remains below 13 nm. Furthermore, both OCNet w/ and w/o domain features achieve lower MAEs compared to the reported SOTA model, demonstrating the effectiveness of our molecular representations pre-trained on conjugated systems. In addition, OCNet w/ domain features exhibits slightly higher accuracy in terms of both MAE and R^2 metrics. This indicates that integration with domain knowledge can enhance the representational capacity of OCNet, further relieving the data scarcity in materials science.

Performance of OCNet on transport task

To demonstrate the performance of OCNet in transport-related tasks, we utilize OCNet's bimolecular representations to predict intermolecular electronic couplings (transfer integrals) in both crystalline and film environments. We first evaluate the performance of OCNet on the OCELOT dimer dataset,[13,46] a diverse dataset containing 438,000 DFT-derived transfer integrals from about 25,000 molecular crystal structures. Inspired by the work of Valeev et

al.,[48] we select domain features including the distance between molecular centroids, the angle between molecular plane normal vectors, and the angle between the centroid-to-centroid vector and the molecular plane normal vectors. Subsequently, we demonstrate the correlation between the predictions of the domain feature-integrated OCNet and the corresponding QM calculation results. Leveraging pre-training on bimolecular conformations and TB level transfer integrals, we achieved accurate predictions of H-H and L-L transfer integrals. in the OCELOT dimer dataset, with MAEs of 0.058 eV and 0.061 eV, respectively, and R^2 values of 0.909 for both cases. We also compare the accuracy of OCNet with reported SOTA model in predicting H-H and L-L transfer integrals (Figure 4c, Tables S5 and S6). We set the accuracy score in Figure 4c as the ratio of the MAE of the SOTA model to the MAE of OCNet. Compared to the prediction accuracy of the SOTA model for H-H and L-L transfer integrals. (MAEs: 0.081 eV and R^2:0.83), our OCNet demonstrates significantly improved performance, achieving a 50% increase in accuracy score. In addition, the performance of OCNet w/o pre-training is inferior to the reported SOTA model (Figure 4d). For instance, in the prediction of H-H transfer integrals, the MAE for OCNet w/o pre-training is approximately 0.12 eV. These results further highlight the superiority of our bimolecular molecular representations, pre-trained on large-scale datasets, in accurately describing electron or hole transfer tasks.

Considering that the functional layers of functional material devices primarily exist in film states, we construct the first DFT-accuracy database (detailed in the methods section) comprising 1.8 OCNet transfer integrals across 5,500 distinct molecular types. We construct the domain features by integrating structural features: the distance between centroids, the angle between plane normal vectors, and the angle between plane normal vectors and the centroid connection vector, and TB-level electronic properties. These electronic properties include overlap integrals, transfer integrals, effective integrals for the HOMO or LUMO orbital, and total effective transfer integrals for all occupied or unoccupied molecular orbitals. Subsequently, we utilize the domain feature-integrated OCNet to predict the transfer inte-

grals database in film environments. As no other models have yet reported similar results, the accuracy score of OCNet is set to 1 (Figure 4c). We further evaluate the correlation between OCNet predictions and corresponding QM values for H-H and L-L transfer integrals(Figure 4b). OCNet demonstrates high accuracy, achieving R^2 values of 0.844 and 0.872 and MAEs of 0.2 eV and 0.204 eV for H-H and L-L transfer integrals, respectively. This marks the first successful realization of a general model for transfer integral prediction in film environments. The accuracy of this general model is comparable to that of reported SOTA models for predicting transfer integrals in crystalline environments, making it suitable for subsequent calculations of molecular mobility in film environments.

III. Conclusion

In summary, we present OCNet, a framework designed to acquire generalized molecular and bimolecular representations for both optoelectronic and transport properties by leveraging SE(3) transformer architectures to automatically extract molecular structural information. To ensure the acquisition of high-quality molecular and bimolecular representations with low data requirements, we pre-train the SE(3) transformer models using a self-constructed, large-scale database containing optoelectronic and electronic coupling properties. For specific downstream tasks, OCNet integrates domain-specific features with general molecular representations, enhancing its expressive power and interpretability. On QM optoelectronic datasets, experimental optoelectronic datasets, and transport-related transfer integral datasets, OCNet significantly outperforms reported state-of-the-art models in prediction accuracy. Additionally, we construct the first large-scale bimolecular transfer integral database, comprising 1.8 million DFT-accuracy data points derived from 55,000 films. This advancement demonstrates OCNet's capability to achieve quantitative predictions on the dataset, laying a foundation for developing generalizable models for transfer integrals in thin-film systems. Consequently, OCNet provides a unified framework for the virtual characterization of

both optoelectronic and transport properties in functional materials, facilitating the establishment of structure-property-function relationships in optoelectronic devices. Furthermore, OCNet's molecular and bimolecular representations have great potential to be transferred to other areas related to conjugated molecular materials, such as photocatalysis and dyes.

Looking forward, by integrating experimental factors, such as temperature, device preparation conditions, and component ratios, into OCNet as domain knowledge features, the scope of OCNet can be expanded from simply predicting material functional properties to evaluating important industrialization-related indicators like device stability and lifespan. The iterative combination of experimental and AI-assisted virtual characterization via OCNet will significantly reduce the time and cost associated with experimental characterization at every stage of molecular design and device preparation. This will undoubtedly accelerate the research and development cycle of functional material devices, driving the progress of this field.

IV. Methods

Details of large-scale molecular and bimolecular database

The large-scale molecular database includes 12M conjugated molecules with structures optimized at the GFN2-xTB level[39] and opto-electronic properties (HOMO-LUMO GAP, absorption energy, absorption transition dipole moment) calculated at the sTDA-GFN2-xTB level.[40] Part of the structures is mined from two open-source databases: an iridium complex database[42] with 0.84 M structures and COMPAS-2x database[43] with 0.5 M cata-condensed hetero-polycyclic aromatic molecules. We generate the remaining structures through aromatic or anti-aromatic ring fusion and fragment assembly.

Firstly, we generate condensed hetero-polycyclic aromatic molecules using 12 kinds of five or six-membered aromatic or anti-aromatic hetero-rings (Figure S1a) as building blocks. To cover the chemical space of condensed hetero-polycyclic aromatic molecules for organic op-

toelectronic applications, our building blocks feature various compositions (including mono- and di-substituted variants with B, N, O, S, and C=O), which are commonly reported in previous studies. Furthermore, when fusing the rings in original condensed hetero-polycyclic aromatic molecules and building blocks (Figure S1b), we allow carbon or hetero atoms to be shared by two or three rings. This approach ensures that part of the generated molecules exhibit multiple resonance features, which are widely used for the preparation of devices with narrow band emission, high luminescence quantum efficiency[49–51] and have not been considered in previous open-source datasets.[43,44,52] To avoid unnecessary resource consumption, we set the maximum ring number to 15, ultimately generating 2 M condensed hetero-polycyclic aromatic molecules.

Besides condensed hetero-polycyclic aromatic molecules and metal-organic complexes, we also create fragment assembly-based aromatic molecules by connecting carbon atoms on the rings that belong to different conjugated molecules(Figure S3). To achieve this, we mine various conjugated molecules from previous publications using the molecular recognition app,[53,54] and construct a fragment library by severing the carbon-carbon single bonds between different rings of those molecules. We also mark the bond-breaking atoms of all fragments and connect hydrogen atoms to them for valence saturation. Finally, we construct 9 M fragment assembly-based aromatic molecules by randomly combining molecular fragments through the automated connection of marked carbon atoms.

We construct the large-scale dimer database consisting of 9.5 million dimer conformations sampled from 100,000 molecular films obtained through MD simulations. These 100,000 molecules are randomly sampled from 1cM molecules with relatively low electron or hole reorganization energies in our molecular database, considering that molecules used for transport applications typically exhibit low reorganization energies.[13,55] Based on these 100,000 molecules, we use Packmol[56] to construct initial boxes with a length of 10 nm, each containing 100 monomers. We then run 4 ns MD simulations under 1 bar and 300 K, using the v-rescale thermostat[57] and the c-rescale barostat.[58] During the MD simulations, we choose

OSCFF,[59] a force field we develop for organic conjugated molecules that is compatible with the GAFF force field, to describe the intermolecular interactions within the film. Our OSCFF completes the missing force field parameters for conjugated molecules and shows accuracy in torsional potential scans comparable to the reference QM method. For detailed specifics, please refer to ref.[59] After 4 ns of MD simulation, the densities of all the film systems have converged. We randomly select 90 dimer conformations with a centroid distance less than 1 nm from the final frame of each system's MD simulation and calculate the HOMO-HOMO and LUMO-LUMO transfer integrals at the GFN1-xTB[41] level.

Computational details of downstream properties database

The geometries of the OCELOT chromophore and the Deep4Chem dataset are converted from SMILES using rdkit[60] and are further optimized using GFN2-xTB. We also select 1.8 M conformations from our bimolecular database. To reduce the computational demands, we limit the maximum number of atoms to 200. Additionally, to increase the proportion of conformations with larger transfer integrals in the dataset, we first randomly sample 0.8 M conformations from the bimolecular database, and then randomly select 1 M conformations from those with HOMO-HOMO or LUMO-LUMO transfer integrals greater than 0.27 eV, calculated at the GFN1-xTB level. The final set with 1.8 million bimolecular conformations contains 5.5k unique molecular types, exhibiting high chemical diversity. We calculate all HOMO-HOMO and LUMO-LUMO transfer integrals for the bimolecular conformations at the PW91/6-31G(d) level[61] using MOMAP[62] and Gaussian16,[63] as the combination of PW91 and 6-31G(d) has demonstrated high accuracy in calculating transfer integrals and carrier mobility.[64,65]

Details of molecular and bimolecular representation

To construct an accurate and general molecular and bimolecular representation for organic functional materials, we employ atomic numbers and pairwise distances as the initial repre-

sentation to encode both atomic and 3D spatial information of the molecular or bimolecular system. Then, we employ the self-attention mechanism in Transformer architecture to couple and update the initial atomic and pair representations, thereby obtaining representations that accurately capture the complex interactions within the molecules or bimolecules. Similar to using the CLS token as a sequence representation aggregator for 1D sequence tasks in the BERT model,[32] we choose the geometric centers of the molecules or bimolecules as the CLS atoms to aggregate atomic features, thereby reflecting the overall structural characteristics of the molecules or bimolecules. We denote the initial atomic representation as:

$$\mathbf{x}^0 = [\mathrm{emb(CLS)}, \mathrm{emb}(Z_0) \ldots \mathrm{emb}(Z_n), \mathrm{emb(PAD)}, \ldots \mathrm{emb(PAD)}]_{n_{\max}+1} \tag{1}$$

where Z_i represents the vocabulary index of the i-th atom in the molecule or bimolecule. all atoms within the molecule or bimolecule are encoded using embedding layer according to their elements, while the first element in eq (1) represents the embedding layer of CLS atom. $n_{\max}$ refers to the maximum number of atoms in a molecule or bimolecule within the database. We employ PAD token to ensure a fixed input size when the number of atoms is less than $n_{\max}$. The initial pair representation is denoted as the molecular or bimolecular distance kernel matrix $\mathbf{P}^0$, where $P_{ij} = \sigma(a_{ij}D_{ij} + b_{ij})$, with a_{ij} and b_{ij} determined by the elemental types of atoms i and j. The L2 distance matrix $\mathbf{D}$ is denoted as:

$$\mathbf{D} = \begin{pmatrix} r_{\mathrm{CLS,CLS}} & r_{\mathrm{CLS},1} & \cdots & r_{\mathrm{CLS},n} & \cdots & 0 \\ r_{1,\mathrm{CLS}} & r_{1,1} & \cdots & r_{1,n} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \cdots & \\ r_{n,\mathrm{CLS}} & r_{d,1} & \cdots & r_{n,n} & \cdots & 0 \end{pmatrix}_{\mathrm{n_{max}+1,n_{max}+1}} \tag{2}$$

If conjugated molecules are in solution environments, we concatenate the initial atomic representations $\mathbf{x}^0_{\mathrm{solu}}$ and $\mathbf{x}^0_{\mathrm{solu}}$ of the solute and solvent molecules, as well as the initial pair representations $\mathbf{P}^0_{\mathrm{solu}}$ and $\mathbf{P}^0_{\mathrm{solu}}$ (Figure S5), to form the initial atomic representations:

$$\mathbf{x}^0 = \left[\text{emb(CLS)}, \text{emb}(Z_0^{\text{solu}}), \ldots, \text{emb}(Z_m^{\text{solu}}), \text{emb}(Z_0^{\text{solv}}), \ldots, \text{emb}(Z_n^{\text{solv}})\right]_{n_{\max}^{\text{solu}}+n_{\max}^{\text{solv}}+1} \quad (3)$$

where Z denotes the vocabulary index of the i-th atom of molecules, and $n_{\max}^{\text{solu}}$ and $n_{\max}^{\text{solv}}$ represents the maximum number of atoms in the solute molecule and solvent molecule. Padding may be applied when the number of atoms in a given system is smaller than M.

and the initial pair representations:

$$\mathbf{P}^0 = \begin{pmatrix} \mathbf{P}_{\text{solu}}^0 & 0 \\ 0 & \mathbf{P}_{\text{solv}}^0 \end{pmatrix} \quad (4)$$

The first element in $\mathbf{x}_0^0$ denotes the initial whole representation of the gas molecule or the solute and the solvent molecule.

Based on the initial atomic and pair representations $\mathbf{x}^0$ and $\mathbf{P}^0$, we update the atomic and pair representations with 15 encoder layers (Figure S6). For the l-th layer, we first construct the weight matrix $\mathbf{W}_Q^l, \mathbf{W}_K^l \in \mathbb{R}^{d_{\text{model}} \times d_k}$ and $\mathbf{W}_V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, to obtain the query matrix $\mathbf{Q}^l = \mathbf{x}^{l-1}\mathbf{W}_Q^l$, the value matrix $\mathbf{V}^l = \mathbf{x}^{l-1}\mathbf{W}_V^l$ and the key matrix $\mathbf{K}^l = \mathbf{x}^{l-1}\mathbf{W}_K^l$. By aggregating the atomic and pair representations from the $l-1$th layer, we obtain the atomic and pair representation for the lth layer. The atomic representation is denoted as:

$$\mathbf{x}^l = \mathbf{x}^{l-1} + (\text{softmax}(\frac{\mathbf{Q}^l {\mathbf{K}^l}^T}{\sqrt{d_k}} + \mathbf{P}^{l-1})\mathbf{V}^l)\mathbf{W}_O^l \quad (5)$$

$$\mathbf{x}^l = \mathbf{x}^l + \text{MLP}(x^l) \quad (6)$$

The MLP represents the multilayer perceptron. We use $\mathbf{W}_O \in \mathbb{R}^{d_v \times d_{\text{model}}}$ to project the output of the attention mechanism to the same dimension as $\mathbf{x}^l$. Similarly, we denote the

pair representation as:

$$\mathbf{P}^l = \mathbf{P}^{l-1} + \frac{\mathbf{Q}^l \mathbf{K}^{l^T}}{\sqrt{d_k}} \tag{7}$$

After being processed by Lth encoder layers and MLP in our MRL model, the initial feature of the CLS atom $\mathbf{x}_0^0$ aggregates the features of atomic and pair representations within the molecule or bimolecule. We denote $\mathbf{x}_0^L$ as $\mathrm{CLS}_{\mathrm{repr}}$, which serves as the overall representation of the molecule or biomolecule. To improve the representational capability of $\mathrm{CLS}_{\mathrm{repr}}$, we pre-train our MRL model on large-scale 3D geometries and TB-level properties. Specifically, we first pretrain the model by predicting masked atoms or reconstructing 3D coordinates using SE(3)-equivariant networks. Then, we fine-tune the model on a large-scale optoelectronic or transfer integral database to further refine its learned representations.

For specific downstream property prediction, we can directly construct a model based on our molecular representation as

$$y = \mathrm{MLP}(\mathrm{CLS}_{\mathrm{repr}}) \tag{8}$$

or integrate domain features (Fea) with the molecular representation to construct the model as

$$y = \mathrm{MLP}(\mathrm{concat}(\mathrm{MLP}(\mathrm{CLS}_{\mathrm{repr}}), \mathrm{MLP}(\mathrm{Fea})) \tag{9}$$

Details of model configuration and training process

We construct both the molecular and bimolecular representations with 15 layers and an embedding dimension of 512, using a Gaussian kernel size of 128. We pre-train the molecular and bimolecular representations on 8 Tesla A100 GPUs, which take approximately 20 days to complete. We use the Adam optimizer with a learning rate of 0.001. During the training, we set the gradient clipping to 1.0, and the training lasted for 8 million steps with 20K warmup steps. We use a batch size of 128 and conducted training over 1000 epochs. The hyperparameters for training the downstream organic optoelectronic properties and transfer integrals are detailed in Table S7.

Acknowledgement

The authors gratefully acknowledge the funding support from the AI for Science Institute, Beijing (AISI) and DP Technology Corporation. The computing resources for this work were provided by the Bohrium Cloud Platform (https://bohrium.dp.tech), which is supported by DP Technology, the Hefei Advanced Computing Center of Sugon, and the High-Performance Computing Platform at Peking University.

Competing Interests

The authors declare no competing financial or non-financial interests.

Supporting Information Available

Comprehensive information regarding the comparison of accuracies among various models, the hyperparameters utilized for training downstream properties, the data distribution within the pre-training database, the methodology for molecule generation, and the architecture of the neural network is presented in the Supporting Information.

Data availability

The large-scale database for pre-training will be available subsequent to the publication of this work.

Code availability

The OCNet code and the pre-training model will be available subsequent to the publication of this work.

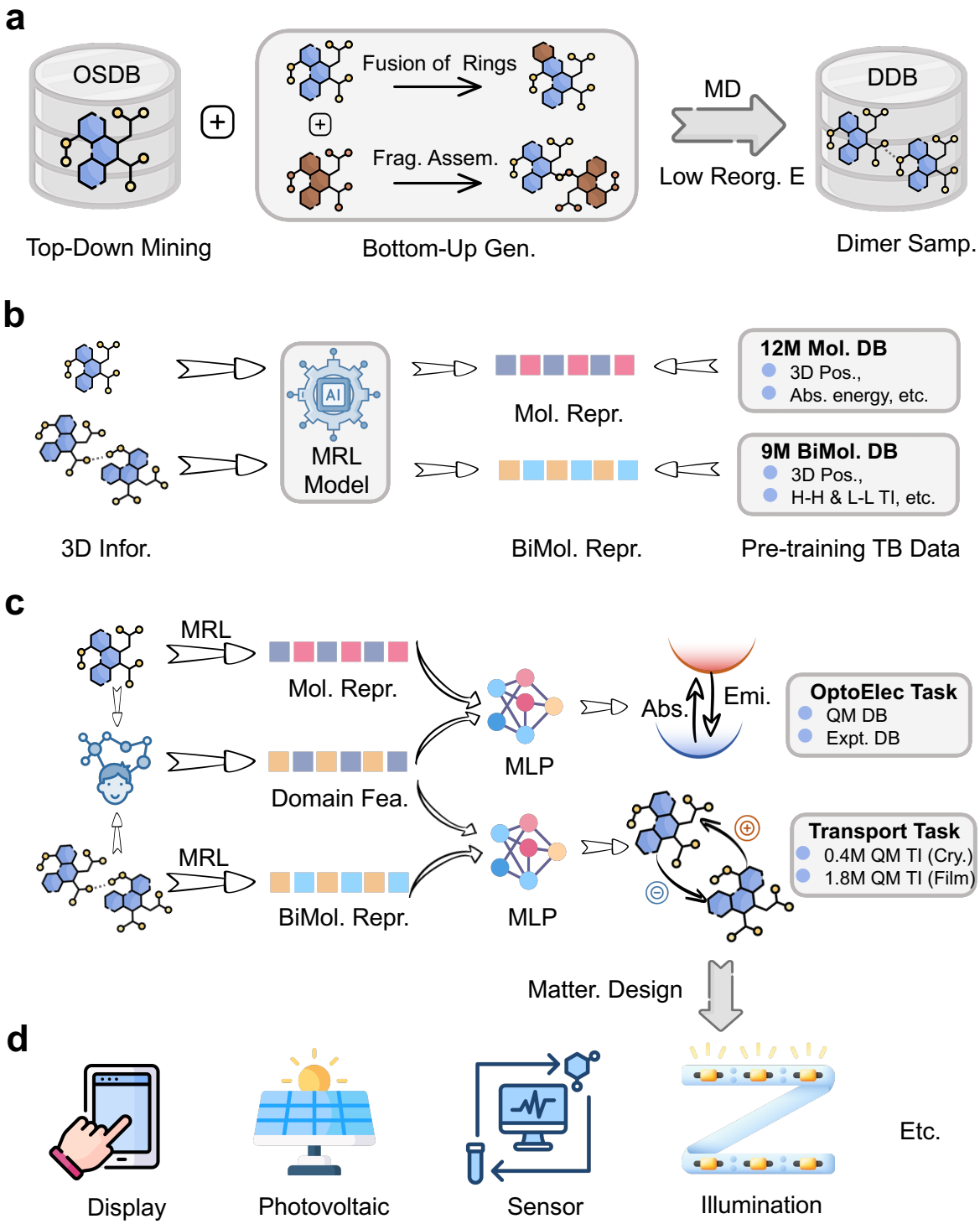

Figure 1: Overview of the OCNet framework: (a) The establishment of two large-scale pre-training databases. We construct the molecular database through mining from open-source databases, generating new molecules using ring fusion and fragment assembly approaches. We also construct the bimolecular database by sampling from 100k molecular films obatined through MD simulations. (b) Developing general molecular and bimolecular representations for conjugated systems. (c) Domain feature integrated molecular representations for downstream property prediction. (d) Potential downstream material design scenarios of OCNet.

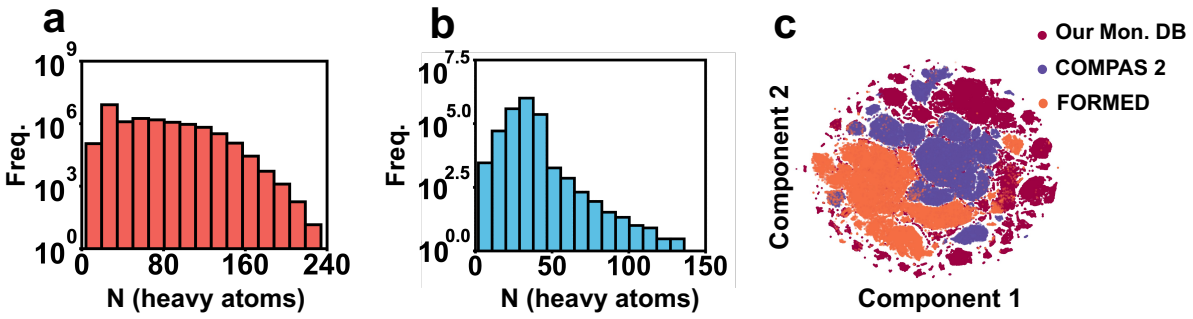

Figure 2: The distribution of heavy atoms across (a) our molecular database, (b) the COMPAS-2x database.[43] (c) Visualization of molecular representations of our molecular database, COMPAS-2x and FORMED[44] using T-SNE analyze.

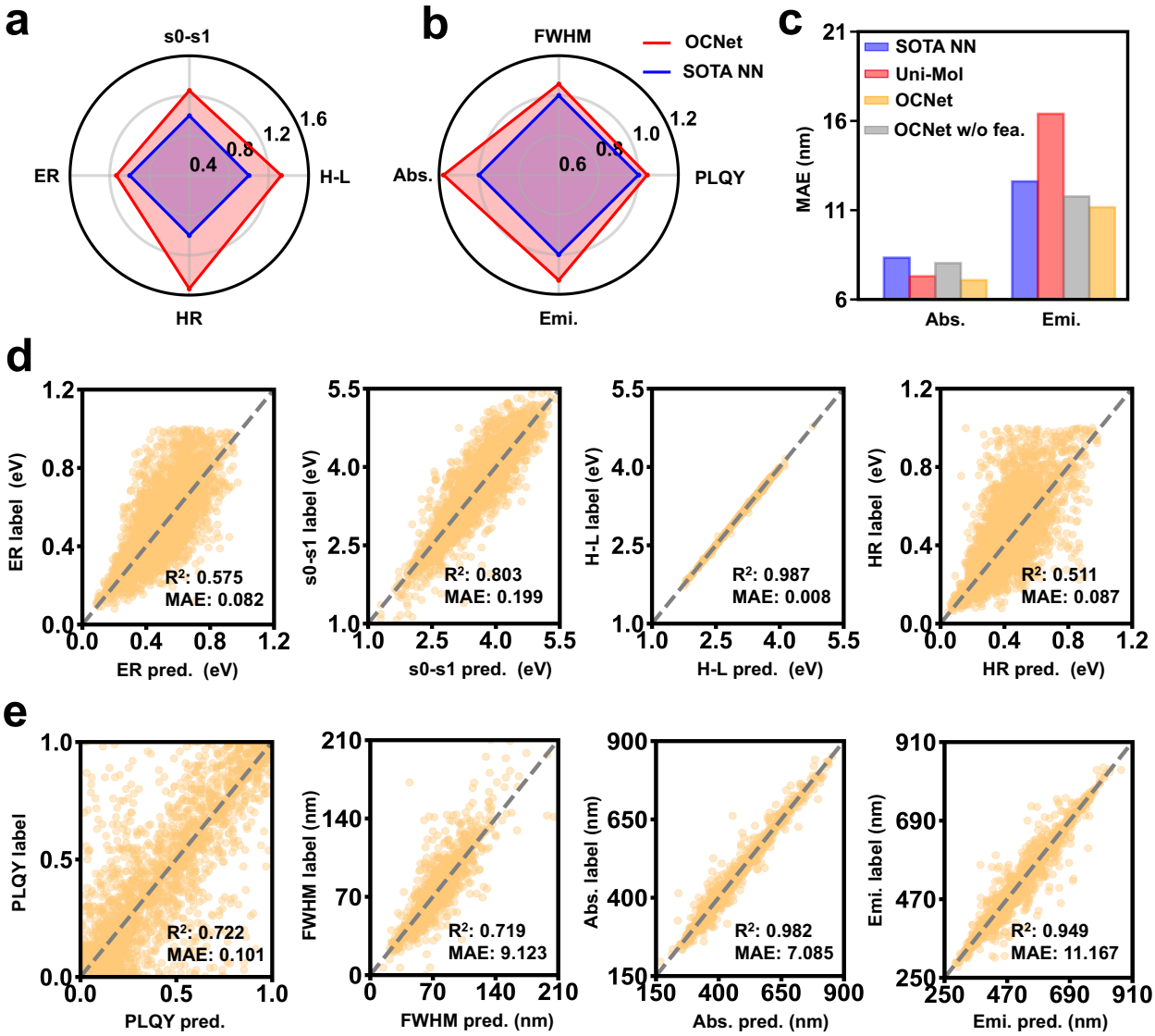

Figure 3: Performance of OCNet on optoelectronic tasks: (a) The accuracy score plots of OCNet (in red) and reported SOTA models (in blue) for predicting QM-calculated properties. (b) The accuracy score plots of OCNet (in red) and reported SOTA models (in blue) for predicting experimental properties. We define the accuracy score of OCNet on the dataset as the ratio of the MAE of reported SOTA model to the MAE of OCNet. (c) The MAEs of reported models, as well as OCNet w/ and w/o pre-training in predicting various optoelectronic properties. (d) The correlation plots between the OCNet predictions and QM-calculated properties(ER, s0-s1, H-L and HR from left to right). (e) The correlation plots between the OCNet predictions and experimental properties (PLQY, FWHM, Abs. and Emi. from left to right).

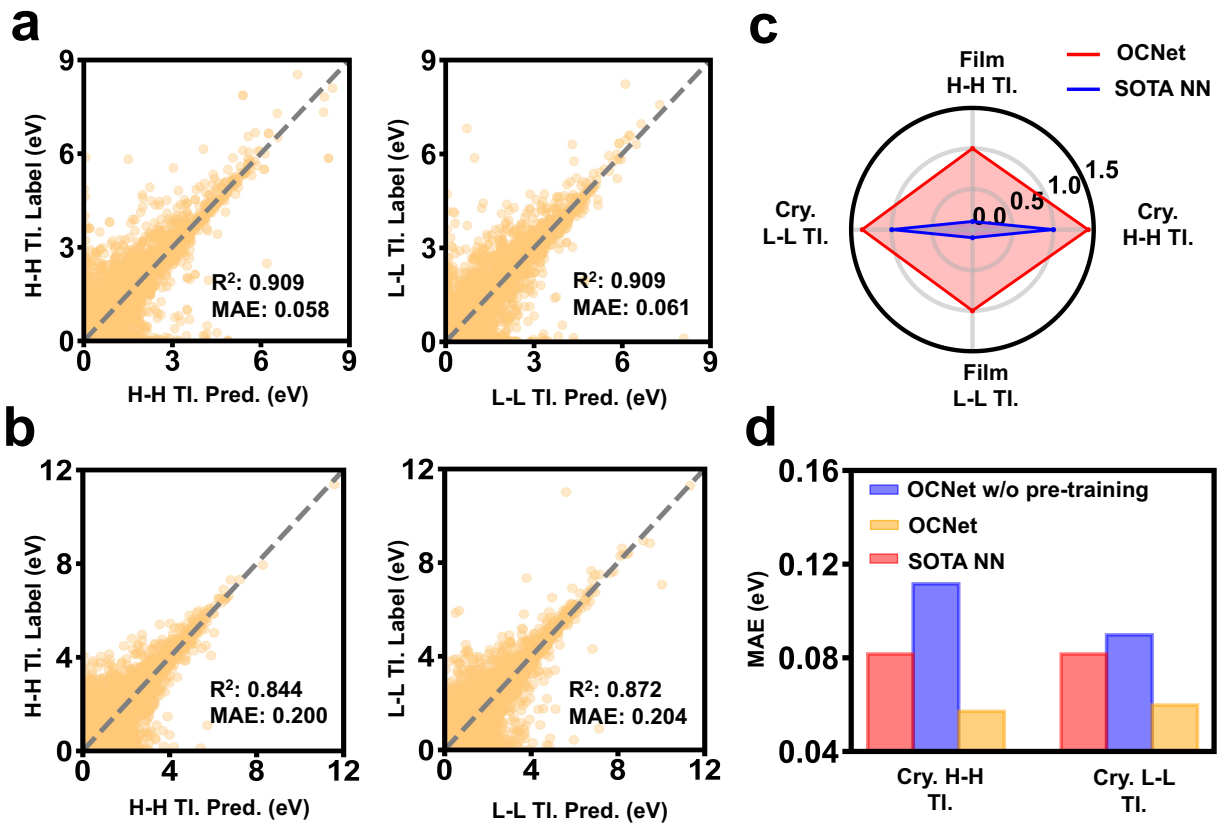

Figure 4: Performance of OCNet on transport tasks: (a) The correlation plots between the OCNet predictions and QM calculated H-H (left) and L-L (right) TI. in molecular crystals. (b) The correlation plots between the OCNet predictions and QM calculated H-H (left) and L-L (right) TI. in molecular films. (c) The accuracy score plots of OCNet (in red) and reported SOTA models (in blue) for predicting transfer integrals. We define the accuracy score of OCNet on the dataset as the ratio of the MAE of reported SOTA model to the MAE of OCNet. As no other models have yet reported similar results, the accuracy score of OCNet for predicting transfer integrals in film environment is set as 1.. (d) The MAEs of reported models, as well as OCNet w/ and w/o pre-training in predicting transfer integrals.

References

(1) Ortega, E. O.; Hosseinian, H.; Meza, I. B. A.; López, M. J. R.; Vera, A. R.; Hosseini, S. *Material characterization techniques and applications*; Springer, 2022.

(2) Zhao, H.; Arneson, C. E.; Fan, D.; Forrest, S. R. Stable blue phosphorescent organic LEDs that use polariton-enhanced Purcell effects. *Nature* **2024**, *626*, 300–305.

(3) Hu, Y.; Wang, J.; Yan, C.; Cheng, P. The multifaceted potential applications of organic photovoltaics. *Nature Reviews Materials* **2022**, *7*, 836–838.

(4) Gkoupidenis, P.; Zhang, Y.; Kleemann, H.; Ling, H.; Santoro, F.; Fabiano, S.; Salleo, A.; van de Burgt, Y. Organic mixed conductors for bioinspired electronics. *Nature Reviews Materials* **2024**, *9*, 134–149.

(5) Wang, J.; Xie, Y.; Chen, K.; Wu, H.; Hodgkiss, J. M.; Zhan, X. Physical insights into non-fullerene organic photovoltaics. *Nature Reviews Physics* **2024**, 1–17.

(6) Ou, Q.; Peng, Q.; Shuai, Z. Computational screen-out strategy for electrically pumped organic laser materials. *Nature Communications* **2020**, *11*, 4485.

(7) Ahmed, S.; Kalita, D. J. Charge transport in isoindigo-dithiophenepyrrole based DA type oligomers: A DFT/TD-DFT study for the fabrication of fullerene-free organic solar cells. *The Journal of chemical physics* **2018**, *149*.

(8) Kümmel, S. Charge-Transfer Excitations: a challenge for time-dependent density functional theory that has been met. *Advanced Energy Materials* **2017**, *7*, 1700440.

(9) Fallon, K. J.; Budden, P.; Salvadori, E.; Ganose, A. M.; Savory, C. N.; Eyre, L.; Dowland, S.; Ai, Q.; Goodlett, S.; Risko, C.; others Exploiting excited-state aromaticity to design highly stable singlet fission materials. *Journal of the American Chemical Society* **2019**, *141*, 13867–13876.

(10) Grotjahn, R.; Maier, T. M.; Michl, J.; Kaupp, M. Development of a TDDFT-based protocol with local hybrid functionals for the screening of potential singlet fission chromophores. *Journal of Chemical Theory and Computation* **2017**, *13*, 4984–4996.

(11) Woo, S.-J.; Kim, Y.-H.; Kim, J.-J. Dihedral angle distribution of thermally activated delayed fluorescence molecules in solids induces dual phosphorescence from charge-transfer and local triplet states. *Chemistry of Materials* **2021**, *33*, 5618–5630.

(12) Troisi, A.; Orlandi, G. Dynamics of the intermolecular transfer integral in crystalline organic semiconductors. *The Journal of Physical Chemistry A* **2006**, *110*, 4065–4070.

(13) Bhat, V.; Ganapathysubramanian, B.; Risko, C. Rapid Estimation of the Intermolecular Electronic Couplings and Charge-Carrier Mobilities of Crystalline Molecular Organic Semiconductors through a Machine Learning Pipeline. *The Journal of Physical Chemistry Letters* **2024**, *15*, 7206–7213.

(14) Tse, S.; So, S.; Yeung, M.; Lo, C.; Wen, S.; Chen, C. The role of charge-transfer integral in determining and engineering the carrier mobilities of 9, 10-di (2-naphthyl) anthracene compounds. *Chemical physics letters* **2006**, *422*, 354–357.

(15) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters* **2012**, *108*, 058301.

(16) Dai, L.; Fu, Y.; Wei, M.; Wang, F.; Tian, B.; Wang, G.; Li, S.; Ding, M. Harnessing electro-descriptors for mechanistic and machine learning analysis of photocatalytic organic reactions. *Journal of the American Chemical Society* **2024**, *146*, 19019–19029.

(17) Chen, Y.; Tian, B.; Cheng, Z.; Li, X.; Huang, M.; Sun, Y.; Liu, S.; Cheng, X.; Li, S.; Ding, M. Electro-Descriptors for the Performance Prediction of Electro-Organic Synthesis. *Angewandte Chemie International Edition* **2021**, *60*, 4199–4207.

(18) Sun, W.; Zheng, Y.; Yang, K.; Zhang, Q.; Shah, A. A.; Wu, Z.; Sun, Y.; Feng, L.; Chen, D.; Xiao, Z.; others Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Science advances* **2019**, *5*, eaay4275.

(19) Sahu, H.; Rao, W.; Troisi, A.; Ma, H. Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. *Advanced Energy Materials* **2018**, *8*, 1801032.

(20) Ye, S.; Hu, W.; Li, X.; Zhang, J.; Zhong, K.; Zhang, G.; Luo, Y.; Mukamel, S.; Jiang, J. A neural network protocol for electronic excitations of N-methylacetamide. *Proceedings of the National Academy of Sciences* **2019**, *116*, 11612–11617.

(21) Ju, C.-W.; Bai, H.; Li, B.; Liu, R. Machine learning enables highly accurate predictions of photophysical properties of organic fluorescent materials: Emission wavelengths and quantum yields. *Journal of Chemical Information and Modeling* **2021**, *61*, 1053–1065.

(22) Kang, B.; Seok, C.; Lee, J. Prediction of molecular electronic transitions using random forests. *Journal of Chemical Information and Modeling* **2020**, *60*, 5984–5994.

(23) Joung, J. F.; Han, M.; Hwang, J.; Jeong, M.; Choi, D. H.; Park, S. Deep learning optical spectroscopy based on experimental database: potential applications to molecular design. *JACS Au* **2021**, *1*, 427–438.

(24) Greenman, K. P.; Green, W. H.; Gómez-Bombarelli, R. Multi-fidelity prediction of molecular optical peaks with deep learning. *Chemical science* **2022**, *13*, 1152–1162.

(25) Sun, M.; Fu, C.; Su, H.; Xiao, R.; Shi, C.; Lu, Z.; Pu, X. Enhancing chemistry-intuitive feature learning to improve prediction performance of optical properties. *Chemical Science* **2024**, *15*, 17533–17546.

(26) Bhat, V.; Sornberger, P.; Pokuri, B. S. S.; Duke, R.; Ganapathysubramanian, B.; Risko, C. Electronic, redox, and optical property prediction of organic π-conjugated molecules through a hierarchy of machine learning approaches. *Chemical Science* **2023**, *14*, 203–213.

(27) Lederer, J.; Kaiser, W.; Mattoni, A.; Gagliardi, A. Machine learning–based charge transport computation for pentacene. *Advanced Theory and Simulations* **2019**, *2*, 1800136.

(28) Rinderle, M.; Kaiser, W.; Mattoni, A.; Gagliardi, A. Machine-learned charge transfer integrals for multiscale simulations in organic thin films. *The Journal of Physical Chemistry C* **2020**, *124*, 17733–17743.

(29) Rinderle, M.; Gagliardi, A. Machine Learning & multiscale simulations: toward fast screening of organic semiconductor materials. 2021 International Conference on Numerical Simulation of Optoelectronic Devices (NUSOD). 2021; pp 1–2.

(30) Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* **2013**, *35*, 1798–1828.

(31) Zhang, D.; Yin, J.; Zhu, X.; Zhang, C. Network representation learning: A survey. *IEEE transactions on Big Data* **2018**, *6*, 3–28.

(32) Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**,

(33) Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; Ke, G. Uni-mol: A universal 3d molecular representation learning framework. **2023**,

(34) Zhu, J.; Xia, Y.; Wu, L.; Xie, S.; Qin, T.; Zhou, W.; Li, H.; Liu, T.-Y. Unified 2d and

3d pre-training of molecular representations. Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. 2022; pp 2626–2636.

(35) Ji, X.; Wang, Z.; Gao, Z.; Zheng, H.; Zhang, L.; Ke, G.; others Uni-Mol2: Exploring Molecular Pretraining Model at Scale. *arXiv preprint arXiv:2406.14969* **2024**,

(36) Zhang, D.; Liu, X.; Zhang, X.; Zhang, C.; Cai, C.; Bi, H.; Du, Y.; Qin, X.; Peng, A.; Huang, J.; others DPA-2: a large atomic model as a multi-task learner. *npj Computational Materials* **2024**, *10*, 293.

(37) Yang, H.; Hu, C.; Zhou, Y.; Liu, X.; Shi, Y.; Li, J.; Li, G.; Chen, Z.; Chen, S.; Zeni, C.; others Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967* **2024**,

(38) Batatia, I.; Benner, P.; Chiang, Y.; Elena, A. M.; Kovács, D. P.; Riebesell, J.; Advincula, X. R.; Asta, M.; Avaylon, M.; Baldwin, W. J.; others A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096* **2023**,

(39) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation* **2019**, *15*, 1652–1671.

(40) Grimme, S.; Bannwarth, C. Ultra-fast computation of electronic spectra for large systems by tight-binding based simplified Tamm-Dancoff approximation (sTDA-xTB). *The Journal of chemical physics* **2016**, *145*.

(41) Grimme, S.; Bannwarth, C.; Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (Z= 1–86). *Journal of chemical theory and computation* **2017**, *13*, 1989–2009.

(42) Cheng, Z.; Liu, J.; Jiang, T.; Chen, M.; Dai, F.; Gao, Z.; Ke, G.; Zhao, Z.; Ou, Q. Automatic Screen-out of Ir (III) Complex Emitters by Combined Machine Learning and Computational Analysis. *Advanced Optical Materials* **2023**, *11*, 2301093.

(43) Mayo Yanes, E.; Chakraborty, S.; Gershoni-Poranne, R. COMPAS-2: a dataset of catacondensed hetero-polycyclic aromatic systems. *Scientific Data* **2024**, *11*, 97.

(44) Blaskovits, J. T.; Laplaza, R.; Vela, S.; Corminboeuf, C. Data-Driven Discovery of Organic Electronic Materials Enabled by Hybrid Top-Down/Bottom-Up Design. *Advanced Materials* **2024**, *36*, 2305602.

(45) Gao, Y.-C.; Yuan, Y.-H.; Huang, S.; Yao, N.; Yu, L.; Chen, Y.-P.; Zhang, Q.; Chen, X. A Knowledge–Data Dual-Driven Framework for Predicting the Molecular Properties of Rechargeable Battery Electrolytes. *Angewandte Chemie International Edition* **2024**, e202416506.

(46) Ai, Q.; Bhat, V.; Ryno, S. M.; Jarolimek, K.; Sornberger, P.; Smith, A.; Haley, M. M.; Anthony, J. E.; Risko, C. OCELOT: An infrastructure for data-driven research to discover and design crystalline organic semiconductors. *The Journal of Chemical Physics* **2021**, *154*.

(47) Joung, J. F.; Han, M.; Jeong, M.; Park, S. Experimental database of optical properties of organic compounds. *Scientific data* **2020**, *7*, 295.

(48) 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. *Journal of the American Chemical Society* **2006**, *128*, 9882–9886.

(49) Wu, X.; Su, B.-K.; Chen, D.-G.; Liu, D.; Wu, C.-C.; Huang, Z.-X.; Lin, T.-C.; Wu, C.-H.; Zhu, M.; Li, E. Y.; others The role of host–guest interactions in organic emitters employing MR-TADF. *Nature Photonics* **2021**, *15*, 780–786.

(50) Madayanad Suresh, S.; Zhang, L.; Hall, D.; Si, C.; Ricci, G.; Matulaitis, T.; Slawin, A. M.; Warriner, S.; Olivier, Y.; Samuel, I. D.; others A Deep-Blue-Emitting Heteroatom-Doped MR-TADF Nonacene for High-Performance Organic Light-Emitting Diodes. *Angewandte Chemie International Edition* **2023**, *62*, e202215522.

(51) Jin, P.; Wei, X.; Yin, B.; Xu, L.; Guo, Y.; Zhang, C. Stepwise Charge/Energy Transfer in MR-TADF Molecule Doped Exciplex for Ultralong Persistent Luminescence Activated with Visible Light. *Advanced Materials* **2024**, 2400158.

(52) Wahab, A.; Pfuderer, L.; Paenurk, E.; Gershoni-Poranne, R. The compas project: A computational database of polycyclic aromatic systems. phase 1: cata-condensed polybenzenoid hydrocarbons. *Journal of Chemical Information and Modeling* **2022**, *62*, 3704–3713.

(53) Xiong, J.; Liu, X.; Li, Z.; Xiao, H.; Wang, G.; Niu, Z.; Fei, C.; Zhong, F.; Wang, G.; Zhang, W.; others αExtractor: a system for automatic extraction of chemical information from biomedical literature. *Science China Life Sciences* **2024**, *67*, 618–621.

(54) Cai, H.; Cai, X.; Chang, J.; Li, S.; Yao, L.; Wang, C.; Gao, Z.; Wang, H.; Li, Y.; Lin, M.; others Sciassess: Benchmarking llm proficiency in scientific literature analysis. *arXiv preprint arXiv:2403.01976* **2024**,

(55) Marcus, R. A. Electron transfer reactions in chemistry: theory and experiment (Nobel lecture). *Angewandte Chemie International Edition in English* **1993**, *32*, 1111–1121.

(56) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *Journal of computational chemistry* **2009**, *30*, 2157–2164.

(57) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of chemical physics* **2007**, *126*.

(58) Bernetti, M.; Bussi, G. Pressure control using stochastic cell rescaling. *The Journal of Chemical Physics* **2020**, *153*.

(59) Zhao, G.; Hu, T.; Wang, H.; Wu, L.; Lu, S.; Yang, F.; Chen, S.; Gao, Z.; Wang, X.; Cheng, Z. Data-Driven Parametrization of All-atom force fields for Organic Semiconductors. **2024**,

(60) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF force field to the RDKit: implementation and validation. *Journal of cheminformatics* **2014**, *6*, 1–4.

(61) Burke, K.; Perdew, J. P.; Wang, Y. *Electronic Density Functional Theory: recent progress and new directions*; Springer, 1998; pp 81–111.

(62) Peng, Q.; Yi, Y.; Shuai, Z.; Shao, J. Toward Quantitative Prediction of molecular fluorescence quantum efficiency: Role of Duschinsky rotation. *Journal of the American Chemical Society* **2007**, *129*, 9333–9339.

(63) Frisch, M. J. et al. Gaussian˜16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(64) Chang, Y.-f.; Lu, Z.-y.; An, L.-j.; Zhang, J.-p. From molecules to materials: molecular and crystal engineering design of organic optoelectronic functional materials for high carrier mobility. *The Journal of Physical Chemistry C* **2012**, *116*, 1195–1199.

(65) Nan, G.; Shi, Q.; Shuai, Z.; Li, Z. Influences of molecular packing on the charge mobility of organic semiconductors: From quantum charge transfer rate theory beyond the first-order perturbation. *Physical Chemistry Chemical Physics* **2011**, *13*, 9736–9746.