SPACIER: A Dataset for Modeling Electrostatic Poisson–Boltzmann Atomic Solvation Potentials

Yongxian Wu
University of California, Irvine
yongxian.wu@uci.edu

Ray Luo
University of California, Irvine
ray.luo@uci.edu

Abstract

Electrostatic solvation free energy is central to biomolecular modeling, yet existing datasets for machine learning are limited in size, resolution, or scope. We introduce SPACIER, a benchmark dataset for Poisson-Boltzmann (PB)-based electrostatics with atomic-level annotations across diverse molecular systems, from small molecules to large protein complexes. Unlike existing solvation datasets, SPACIER emphasizes atomic precision in water while covering a broad spectrum of system sizes. Grounded in the PB equation, SPACIER enables evaluation of both molecular learning methods and neural PDE solvers under standardized preprocessing, metrics, and loss functions. We further propose a charge-weighted regression objective that improves training stability by mitigating variance in atomic potentials. Baseline experiments with U-Net, Fourier Neural Operator, and graph neural network demonstrate competitive accuracy and scalability, but also reveal limitations in robustness and generalization. By framing solvation modeling as a physically grounded dataset task, SPACIER provides a foundation for advancing machine learning and PDE-based methods in biomolecular electrostatics [1].

1 Introduction

Electrostatic interactions are central to molecular behavior, and the Poisson–Boltzmann (PB) equation provides a rigorous framework for modeling them in ionic solvents [Baker et al., 2001]. As a cornerstone of biomolecular electrostatics, the PB model enables solvation energy estimation, which is critical for applications such as protein–ligand binding [Sharp and Honig, 1990], enzyme catalysis [Page and Jencks, 1971], and DNA transcription regulation [Record Jr et al., 1976].

Despite its strong physical grounding, the PB equation remains challenging to solve efficiently for large biomolecular systems. Numerical methods such as finite difference [Nicholls and Honig, 1991, Bruccoleri et al., 1997, Grant et al., 2001] and finite element [Holst et al., 2012, Chen et al., 2007] are standard and have been implemented in software including AMBER PBSA [Cai et al., 2009], Delphi [Li et al., 2013], APBS [Baker et al., 2001], MIBPB [Zhou et al., 2006], and CHARMM PBEQ [Jo et al., 2008]. However, their computational cost limits scalability. Approximate approaches such as the Generalized Born (GB) model [Bashford and Case, 2000] improve efficiency by simplifying solvation effects, but sacrifice accuracy, especially near highly charged or complex regions. Thus, accurate and efficient solvation free energy estimation remains a bottleneck for molecular dynamics, high-throughput screening, and interactive modeling [Lu et al., 2008].

Recent advances in machine learning, particularly geometric deep learning Bronstein et al. [2021], have led to fast surrogates for PDE solvers Li et al. [2020], improved modeling of molecular properties [Walters and Barzilay, 2020, Wu and Luo, 2025, Wu et al., 2025], and data-driven force fields [Chmiela et al., 2017]. While these approaches promise to accelerate scientific discovery, they rely on large-scale datasets to train modern architectures such as graph neural networks (GNNs).

[1] The SPACIER dataset is available at doi.org/10.5281/zenodo.15867553, and the source code for reproducing our experiments is accessible at github.com/yxwu21/PBGNN.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

Molecular machine learning has produced a range of such datasets. Early benchmarks like QM9 Ramakrishnan et al. [2014] and MD-17 Chmiela et al. [2017] are limited to small molecules (less than 50 atoms) and few atom types. The recently released Open Molecules 2025 (OMol25) dataset [Levine et al., 2025] expands coverage to 100 million DFT calculations for molecules with up to 350 atoms. However, existing datasets emphasize atomic simulations rather than solvation free energy prediction. Several specialized datasets target solvation free energy. QM9-Solvation Wu et al. [2018] extends QM9 with energies in five solvents, while FreeSolv Mobley and Guthrie [2014] provides experimental measurements for 643 neutral molecules. CombiSolv-QM Vermeire and Green [2021] further expands solvent diversity by considering 284 solvents with 11,029 solutes. Despite their breadth, these datasets focus on small molecules and report only total solvation free energies, limiting scalability to large biomolecular systems and precluding atomic-level resolution, which is essential for accurate simulations.

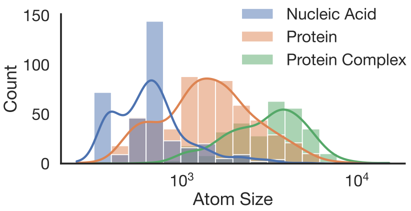


Figure 1: Histogram of atom size distribution in the SPACIER-M.

We introduce SPACIER, a dataset for PB-based electrostatic solvation energy prediction with atomic potential annotations across diverse molecular systems. Unlike prior datasets that emphasize solvent variety, SPACIER focuses on water while spanning small molecules, nucleic acids, proteins, and complexes (1,222 systems, 268 to 140,000 atoms). This design poses unique challenges for machine learning, requiring accurate modeling of short- and long-range interactions. Grounded in the PB equation, SPACIER also serves as a benchmark for neural PDE solvers on large-scale grids with partially observable solutions. We show that SPACIER can be modeled by U-Net [Ronneberger et al., 2015], Fourier Neural Operator [Li et al., 2020], and geometric GNN [Schütt et al., 2017], using standardized losses, metrics, and preprocessing. A charge-weighted regression objective improves training stability, and baseline experiments highlight strong accuracy and scalability but limited robustness and generalizability. By framing solvation free energy prediction as a physically grounded dataset task, SPACIER establishes a foundation for systematic advances at the intersection of machine learning and PDE modeling, with all code, data, and models released openly.

2 SPACIER: Electrostatic PB-Based Atomic Solvation Potentials Prediction

2.1 Preliminary: Solvation Free Energy and Electrostatic Solvation Energy

Solvation free energy (ΔG_{solv}) is the change in Gibbs free energy when transferring a solute from vacuum to solvent. It comprises electrostatic and non-electrostatic components:

$$\Delta G_{\text{solv}} = \Delta G_{\text{elec}} + \Delta G_{\text{non-elec}}. \tag{1}$$

For charged or polar molecules, the electrostatic term (ΔG_{elec}) dominates [Honig and Nicholls, 1995]. Since polar molecules prevail in chemistry and biology, SPACIER focuses on electrostatic solvation energy as the principal contribution in such systems.

Poisson-Boltzmann Equation Given a molecule with N_a atoms, represented as a set $\mathcal{M} = \{\mathbf{a}_1, \ldots, \mathbf{a}_{N_a}\}$, where each atom $\mathbf{a}_i = (\mathbf{r}_i, q_i)$ consists of its three-dimensional spatial position $\mathbf{r}_i$ and charge q_i, the PB equation is widely used to compute the electrostatic solvation energy of biomolecules immersed in ionic solvents [Baker et al., 2001, Fogolari et al., 2002]. The PB equation describes the spatial distribution of the electrostatic potential $\phi(\mathbf{r})$ over two domains: the interior solute region Ω_{int} and the surrounding solvent region Ω_{ext}, separated by a dielectric interface Γ:

$$\nabla \cdot (\epsilon(\mathbf{r}) \nabla \phi(\mathbf{r})) = -\rho(\mathbf{r}) - \lambda(\mathbf{r}) \sum_i n_i q_i \exp\left[-\frac{q_i \phi(\mathbf{r})}{kT}\right], \tag{2}$$

where $\phi(\mathbf{r})$ represents the electrostatic potential at a given field position $\mathbf{r}$, $\epsilon(\mathbf{r})$ is the spatially varying dielectric constant, and $\rho(\mathbf{r})$ is the charge density. The second term on the right hand side represents the mobile ion charge density in the solvent, where $\lambda(\mathbf{r})$ is the masking function for the Stern layer, n_i is the number density of ion type i, q_i is the charge of ion type i, k is the Boltzmann constant, and T is the temperature. The dielectric constant is defined as $\epsilon(\mathbf{r}) = \epsilon_1$ for $\mathbf{r} \in \Omega_{\text{int}}$ and $\epsilon(\mathbf{r}) = \epsilon_2$ for $\mathbf{r} \in \Omega_{\text{ext}}$. Classical solvers such as finite difference and finite element methods [Nicholls and Honig, 1991, Holst et al., 2012] yield accurate PB solutions but are computationally costly for large biomolecules, and are implemented in packages including AMBER PBSA [Cai et al., 2009], Delphi [Li et al., 2013], and APBS [Baker et al., 2001].

Electrostatic Solvation Energy By solving the PB equation in (2), the electrostatic solvation free energy can be computed as [Cai et al., 2009]: $\Delta G_{\text{elec}} = \frac{1}{2}\sum_{i=1}^{N_a} q_i \phi_{\text{rxn}}(\mathbf{r}_i)$, where $\phi_{\text{rxn}}(\mathbf{r}_i)$ denotes the atomic solvation potential (i.e., the reaction field potential) at the position of atom i.

2.2 SPACIER Dataset and Machine Learning Tasks

SPACIER The SPACIER centers on water, the most widely studied solvent, while spanning molecular systems from small molecules to protein complexes. It consists of two subtasks: SPACIER-M and SPACIER-S. The SPACIER-M task contains 1,222 biomolecular systems, including proteins, nucleic acids, and protein complexes, with sizes ranging from 268 to over 140,000 atoms. This dataset presents a significant challenge for developing efficient machine learning methods capable of extracting physically meaningful interactions from large node sets and solving large-scale PDEs. In addition, SPACIER includes a complementary subtask, SPACIER-S, which is designed to evaluate the generalization of machine learning methods to small-molecule electrostatics. The SPACIER-S dataset comprises 812 drug-like molecules ranging from 3 to 62 atoms, providing a convenient setting for researchers to conduct rapid prototyping and preliminary evaluations.

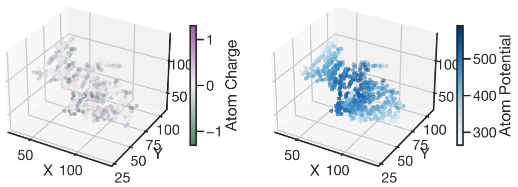


Figure 2: A nucleic acid example (1A2E) from SPACIER-M.

To calculate the electrostatic solvation energy, we employed the AMBER PBSA solver [Case et al., 2023a], which is based on finite difference methods. All molecules were preprocessed using standard topology radii (`mbondi`), a probe radius of 1.4 Å, and a grid spacing of 0.35 Å. The computed atomic solvation potentials were stored, and each molecule was represented in coordinate sparse format together with additional metadata, including molecular identity, atomic charges, spatial coordinates, atomic radii, grid size and spacing used in PDE solving, and per-atom potentials. The SPACIER-M and SPACIER-S datasets were then split into training and test sets using an 80/20 ratio, ensuring no overlap in molecular identities.

Atomic Solvation Potentials Prediction Given a molecule with N_a atoms, a machine learning model f_θ is trained to map the molecular representation $\mathcal{M}$ to a set of predicted atomic potentials $\hat{y}_i$, such that the predicted electrostatic solvation energy $\Delta\hat{G}_{\text{elec}}$ is given by $\Delta\hat{G}_{\text{elec}} = \frac{1}{2}\sum_{i=1}^{N_a} q_i \hat{y}_i$. In our experiments, we demonstrate that this prediction task can be addressed from three complementary perspectives: (1) voxel-based U-Nets [Ronneberger et al., 2015], which predict potentials over a discretized 3D grid; (2) neural operators [Li et al., 2020], which directly approximate the solution of the PB PDE; and (3) geometric graph neural networks [Schütt et al., 2017], which incorporate molecular structure while preserving translation and rotation invariance. To evaluate their efficiency and accuracy in approximating PB energies, all models are trained using supervised regression on atomic solvation potentials. In practice, we observed that direct optimization of atomic potentials using standard mean squared error (MSE) often suffers from instability due to the large variance in $\phi_{\text{rxn}}(\mathbf{r}_i)$. To address this issue, we propose a charge-weighted mean squared error (CMSE) loss function that incorporates atomic charges to emphasize physically meaningful contributions $\mathcal{L}_{\text{CMSE}} = \frac{1}{N_a}\sum_{i=1}^{N_a} |q_i| \left(\hat{y}_i - \phi_{\text{rxn}}(\mathbf{r}_i)\right)^2$. This loss function improves both optimization stability and physical fidelity of the predicted energies, as demonstrated in our experiments.

3 Experiments and Results

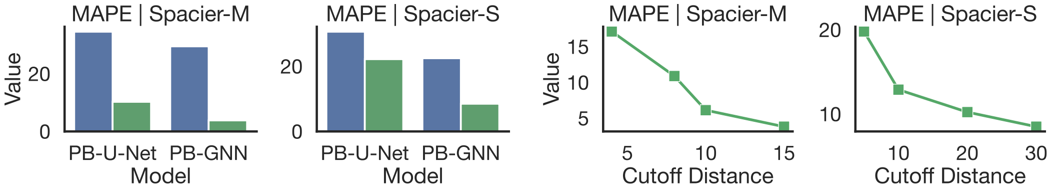
Three models, U-Net [Ronneberger et al., 2015], Fourier Neural Operator (FNO) [Li et al., 2020], and geometric GNN [Schütt et al., 2017], are adopted in the experiments to evaluate performance on the SPACIER. In addition, we include the GB model with `igb=1` from the PMEMD program [Case et al., 2023b] to compare the machine learning models against a classical approximation method.

Table 1: Performance comparison of electrostatic solvation energy. Results are shown for two subtasks: SPACIER-M (left) and SPACIER-S (right).

Method	# Param.	$R^2\uparrow$	MAE $\downarrow$	MAPE $\downarrow$	Method	# Param.	$R^2\uparrow$	MAE $\downarrow$	MAPE $\downarrow$
GB Model	-	0.990	**123.350**	4.190	GB Model	-	0.965	0.510	13.680
PB-U-Net	2.0M	0.931	481.947	10.268	PB-U-Net	2.0M	0.915	1.204	22.187
PB-FNO	16.7M	0.939	513.808	11.649	PB-FNO	16.7M	0.901	1.332	23.359
PB-GNN	0.5M	**0.995**	151.540	**3.822**	PB-GNN	0.5M	**0.968**	**0.475**	**8.503**

Implementation Details To enable modeling of large biomolecules, we construct cutoff-based molecular graphs using approximate nearest neighbor search [Malkov and Yashunin, 2018], thereby avoiding the quadratic complexity of all-pairs distance computations. For U-Net and FNO, we utilize 3D atomic grid representations with width, height, and depth of 128 as input features. For GNN-based models, atomic features are encoded using sinusoidal charge embeddings, and message passing captures both local and semi-local spatial interactions. We report three standard regression metrics: the coefficient of determination (R^2), mean absolute error (MAE), and mean absolute percentage error (MAPE) between the predicted electrostatic solvation energy $\Delta\hat{G}_{\text{elec}}$ and ground truth electrostatic solvation free energy ΔG_{elec}.

Accuracy and Generalization The model performances are summarized in Table 1. On SPACIER-M, PB-GNN achieves the best performance among all models, with MAPE $= 3.82\%$ relative to the reference PB energies. Comparable performance is observed on SPACIER-S, with MAPE $= 8.50\%$. These results highlight the challenges of applying convolutional methods (PB-U-Net) and neural operators (PB-FNO) to large-scale grid spaces, particularly under partial grid observations. Moreover, the performances in terms of MAE and MAPE on SPACIER-S still indicate substantial room for improvement, which may be addressed by incorporating richer environmental information. In comparison to GB models, on SPACIER-M, PB-GNN achieves a lower MAPE (3.82%) than GB (4.19%). On SPACIER-S, the improvement of the machine learning model is even more pronounced: PB-GNN attains a MAPE of 8.50%, compared to 13.68% for the GB model. These results demonstrate that machine learning models are capable of accurately predicting electrostatic solvation energies.



(a) Comparison of MSE (blue) and CMSE (green).

(b) Performance under different cutoff distances.

Figure 3: Analysis of the impact of objectives and graph construction choices on performance.

Importance of Objective Function and Environmental Context Training with CMSE yields consistent improvements over MSE. As shown in Figure 3a, on SPACIER-M, CMSE reduces MAPE from 29.5% to 3.82%, and on SPACIER-S, it reduces MAPE from 22.47% to 8.50%. These results validate CMSE as a physically meaningful training objective for approximating electrostatic solvation energies. We further analyzed the effect of cutoff distance on PB-GNN, as shown in Figure 3b. PB-GNN consistently improves performance on SPACIER with larger cutoff distances. However, increasing the cutoff distance substantially raises the computational cost of modeling atomic interactions, posing challenges for large-scale GNN training.

4 Conclusion

We introduced SPACIER, a benchmark dataset for PB-based electrostatics with atomic-level annotations across diverse molecular systems. Through systematic evaluations with U-Net, FNO, and GNN, we demonstrated both the potential and current limitations of machine learning approaches for solvation modeling. Our results highlight the effectiveness of charge-weighted objectives and the importance of structural context in achieving physically faithful predictions. By framing electrostatic solvation energy prediction as a standardized dataset task, SPACIER establishes a foundation for advancing learning-based and PDE-inspired methods in biomolecular electrostatics.

References

Nathan A Baker, David Sept, Simpson Joseph, Michael J Holst, and J Andrew McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, 98(18):10037–10041, 2001.

Kim A Sharp and Barry Honig. Electrostatic interactions in macromolecules: theory and applications. *Annual review of biophysics and biophysical chemistry*, 19(1):301–332, 1990.

Michael I Page and William P Jencks. Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect. *Proceedings of the National Academy of Sciences*, 68(8):1678–1683, 1971.

M Thomas Record Jr, Timothy M Lohman, and Pieter De Haseth. Ion effects on ligand-nucleic acid interactions. *Journal of molecular biology*, 107(2):145–158, 1976.

Anthony Nicholls and Barry Honig. A rapid finite difference algorithm, utilizing successive over-relaxation to solve the poisson–boltzmann equation. *Journal of computational chemistry*, 12(4): 435–445, 1991.

Robert E Bruccoleri, Jiri Novotny, Malcolm E Davis, and Kim A Sharp. Finite difference poisson-boltzmann electrostatic calculations: Increased accuracy achieved by harmonic dielectric smoothing and charge antialiasing. *Journal of computational chemistry*, 18(2):268–276, 1997.

J Andrew Grant, Barry T Pickup, and Anthony Nicholls. A smooth permittivity function for poisson–boltzmann solvation methods. *Journal of computational chemistry*, 22(6):608–640, 2001.

Michael Holst, James Andrew Mccammon, Zeyun Yu, YC Zhou, and Yunrong Zhu. Adaptive finite element modeling techniques for the poisson-boltzmann equation. *Communications in computational physics*, 11(1):179–214, 2012.

Long Chen, Michael J Holst, and Jinchao Xu. The finite element approximation of the nonlinear poisson–boltzmann equation. *SIAM journal on numerical analysis*, 45(6):2298–2320, 2007.

Qin Cai, Jun Wang, Hong-Kai Zhao, and Ray Luo. On removal of charge singularity in poisson–boltzmann equation. *The Journal of Chemical Physics*, 130(14):145101, 04 2009. ISSN 0021-9606. doi: 10.1063/1.3099708.

Lin Li, Chuan Li, Zhe Zhang, and Emil Alexov. On the dielectric "constant" of proteins: smooth dielectric function for macromolecular modeling and its implementation in delphi. *Journal of chemical theory and computation*, 9(4):2126–2136, 2013.

YC Zhou, Shan Zhao, Michael Feig, and Guo-Wei Wei. High order matched interface and boundary method for elliptic equations with discontinuous coefficients and singular sources. *Journal of Computational Physics*, 213(1):1–30, 2006.

Sunhwan Jo, Miklos Vargyas, Judit Vasko-Szedlar, Benoît Roux, and Wonpil Im. Pbeq-solver for online visualization of electrostatic potential of biomolecules. *Nucleic acids research*, 36(suppl_2): W270–W275, 2008.

Donald Bashford and David A Case. Generalized born models of macromolecular solvation effects. *Annual review of physical chemistry*, 51(1):129–152, 2000.

BZ Lu, YC Zhou, MJ Holst, and JA McCammon. Recent progress in numerical methods for the poisson-boltzmann equation in biophysical applications. *Commun Comput Phys*, 3(5):973–1009, 2008.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

W Patrick Walters and Regina Barzilay. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of chemical research*, 54(2):263–270, 2020.

Yongxian Wu and Ray Luo. Grid-context convolutional model for efficient molecular surface construction from point clouds. *Journal of Chemical Theory and Computation*, 21(15):7648–7661, 2025.

Yongxian Wu, Qiang Zhu, and Ray Luo. End-to-end modeling of reaction field energy using data-driven geometric graph neural networks. *Journal of Chemical Theory and Computation*, 21(19): 9710–9725, 2025.

Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, 2014.

Daniel S Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G Taylor, Muhammad R Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, et al. The open molecules 2025 (omol25) dataset, evaluations, and models. *arXiv preprint arXiv:2505.08762*, 2025.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

David L Mobley and J Peter Guthrie. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of Computer-Aided Molecular Design*, 28(7):711–720, 2014.

Frederik H Vermeire and William H Green. Combisolv-qm and combisolv-exp-extended: datasets for computational solubility prediction and high-throughput solvent screening. *Journal of Chemical Information and Modeling*, 61(12):5971–5987, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Barry Honig and Anthony Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268 (5214):1144–1149, 1995.

Federico Fogolari, Alessandro Brigo, and Henriette Molinari. The poisson–boltzmann equation for biomolecular electrostatics: a tool for structural biology. *Journal of Molecular Recognition*, 15(6): 377–392, 2002.

David A Case, Hasan Metin Aktulga, Kellon Belfon, David S Cerutti, G Andrés Cisneros, Vinícius Wilian D Cruzeiro, Negin Forouzesh, Timothy J Giese, Andreas W Gootz, Holger Gohlke, et al. Ambertools. *Journal of chemical information and modeling*, 63(20):6183–6191, 2023a.

David A Case, Nikolai R Skrynnikov, Thomas E Cheatham III, Oleg Mikhailovskii, Carlos Simmerling, Yi Xue, Adrian Roitberg, Yi Xue, Adrian Roitberg, Sergei A Izmailov, Kenneth M Merz, Koushik Kasavajhala, et al. *AMBER 23 Reference Manual*. University of California, 2023b.

Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018.