Interaction specificity of clustered protocadherins inferred from sequence covariation and structural analysis

John M. Nicoludis[a,b,1], Anna G. Green[c,1], Sanket Walujkar[d], Elizabeth J. May[a], Marcos Sotomayor[d], Debora S. Marks[c,2], and Rachelle Gaudet[a,2]

[a]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138; [b]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138; [c]Department of Systems Biology, Harvard Medical School, Boston, MA 02115; and [d]Department of Chemistry and Biochemistry, The Ohio State University, Columbus, OH 43210

Edited by William I. Weis, Stanford University School of Medicine, Stanford, CA, and approved July 30, 2019 (received for review December 17, 2018)

Clustered protocadherins, a large family of paralogous proteins that play important roles in neuronal development, provide an important case study of interaction specificity in a large eukaryotic protein family. A mammalian genome has more than 50 clustered protocadherin isoforms, which have remarkable homophilic specificity for interactions between cellular surfaces. A large antiparallel dimer interface formed by the first 4 extracellular cadherin (EC) domains controls this interaction. To understand how specificity is achieved between the numerous paralogs, we used a combination of structural and computational approaches. Molecular dynamics simulations revealed that individual EC interactions are weak and undergo binding and unbinding events, but together they form a stable complex through polyvalency. Strongly evolutionarily coupled residue pairs interacted more frequently in our simulations, suggesting that sequence coevolution can inform the frequency of interaction and biochemical nature of a residue interaction. With these simulations and sequence coevolution, we generated a statistical model of interaction energy for the clustered protocadherin family that measures the contributions of all amino acid pairs at the interface. Our interaction energy model assesses specificity for all possible pairs of isoforms, recapitulating known pairings and predicting the effects of experimental changes in isoform specificity that are consistent with literature results. Our results show that sequence coevolution can be used to understand specificity determinants in a protein family and prioritize interface amino acid substitutions to reprogram specific protein–protein interactions.

clustered protocadherins | protein–protein interactions | sequence covariation | molecular dynamics | polyvalency

Clustered protocadherins (Pcdhs) are a large protein family that play roles in vertebrate nervous system development, including neuronal survival, axon targeting, neuronal arborization, and dendritic self-avoidance (1–9). Dendritic self-avoidance, wherein dendritic arbors are pruned when 2 dendrites from the same neuron come in contact, is mediated by formation of a clustered Pcdh assembly between 2 dendrites (5, 10). This assembly consists of specific homodimers formed in *trans* across 2 cellular membranes, engaging the first 4 extracellular cadherin (EC) repeat domains in an antiparallel arrangement (Fig. 1*A*). In addition to these *trans* interfaces, ECs 5 and 6 interact between protocadherins on the same neuron (*cis*), resulting in a zipper-like lattice (11). The *trans* homodimers are highly specific such that no cross interactions are observed in even the most similar isoforms (12–14). The *trans* EC1-4 interaction is also found in nonclustered Pcdhs (15–17), indicating the importance of this adhesive interface in cognitive function. Given the many isoforms per vertebrate genome, including 53 in the human genome, we sought to understand how specificity is achieved in this large interface.

Structures of Pcdh *trans* dimers (11, 15, 18–20) have revealed idiosyncratic characteristics of individual dimer structures, like the lack of EC1/EC4 interaction in the PcdhγA1 and PcdhγA8 structures (20) and the small EC2/EC3 interface in PcdhγB3 (15). Based on the variety of interfaces found in the existing crystal structures (15, 18–20), it is possible that every isoform achieves specificity by adopting a different static interface conformation, or that isoforms sample a distribution of conformations, with different combinations of interface residues determining preference for self-interaction. Understanding Pcdh interaction specificity requires disentangling these scenarios by considering both interface conformations and residue–residue interaction preferences.

Prior computational work has sought to understand the evolution of specificity of the Pcdh *trans* interaction, finding positive selection on the *trans* interface (21), and suggesting that the EC2/EC3 interface plays a greater role in specificity between closely related protocadherins (10, 15). However, this computational framework could not analyze residue dependencies at the interface, or be used to predict specificities for new mutations or combinations of protocadherins. Recent computational methods based on residue coevolution have proven useful for understanding

Significance

The more than 50 clustered protocadherin isoforms impart neuronal identity to prevent self-synapses. This is accomplished via highly specific homodimers of the first 4 extracellular cadherin (EC) domains between neuronal processes. Using molecular dynamics, we found that individual EC domain interactions are weak and dynamic, suggesting the complex is held together by polyvalency. Using evolutionary couplings—a computational technique that identifies interacting amino acids in a protein—we found that interacting residue pairs with strong coupling scores were in contact more frequently and across more simulations than weak pairs. We built a statistical model of specificity that predicts the relative likelihood that any 2 isoforms interact and identify residues contributing strongly to specificity to reprogram protein–protein interactions.

Author contributions: J.M.N., A.G.G., M.S., D.S.M., and R.G. designed research; J.M.N., A.G.G., and S.W. performed research; J.M.N., A.G.G., and E.J.M. contributed new reagents/analytic tools; J.M.N., A.G.G., and S.W. analyzed data; and J.M.N., A.G.G., S.W., M.S., D.S.M., and R.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The structures have been deposited in the Protein Data Bank, www.wwpdb.org (PcdhγB3 EC1-4 no HEPES [PDB ID 6MEQ] and PcdhγB3 EC1-4 less HEPES [PDB ID 6MER]), and the corresponding raw diffraction images have been deposited to the SBGridDB (datasets 602 and 603, respectively).

[1]J.M.N. and A.G.G. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: debbie@hms.harvard.edu or gaudet@mcb.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1821063116/-/DCSupplemental.

Published online August 20, 2019.

the structure and function of protein complexes (22, 23). These methods use undirected graphical models of protein sequences to find statistical dependencies between pairs of residues (24–26), successfully predicting correct protein–protein interaction pairings for families with many paralogs (27, 28), informing specificity reprogramming experiments (29, 30) and predicting the effect of mutations on protein function (31). Therefore, these generative models of residue dependencies may allow for characterization of residues important to specificity in the clustered Pcdh family, and prediction of interaction probability of all possible isoform pairs.

To better understand the structural and dynamical determinants of Pcdh interaction specificity, we use molecular dynamics (MD) simulations and crystal structures to show that isoforms adopt a range of conformations and to identify interacting residue pairs. We build a statistical model using evolutionary couplings (24, 25) to analyze specificity of all possible *trans* isoform interactions and infer which domains are important for interaction and specificity. This work provides insight into the molecular origins of specificity within the clustered Pcdh family and demonstrates that models based on sequence coevolution and MD simulations can be used to guide reprogramming of protein–protein interaction specificity.

Results

Structures and MD Show That Pcdh *trans* Interfaces Sample a Distribution of Conformations. The *trans* interface of clustered Pcdhs is specific for self-interaction (12, 13) (Fig. 1*A*). Clustered Pcdh paralogs share only 40% residue identity at their interfaces (Fig. 1*C*). Crystal structures of many Pcdh isoform dimers (15, 18–20) revealed idiosyncratic features of individual isoforms, raising the possibility that different isoforms adopt different conformations in vivo, resulting in molecular specificity. For example, PcdhγA1 (in 1 of 2 dimers) and PcdhγA8 (20) lack EC1/EC4 contacts, and PcdhγB3 EC2-3 has a surprisingly small interface (15). Pairs of different isoform static structures share a mean of 32% of their interface residue contacts, with only 14 contact pairs shared between all 4 analyzed isoforms (Fig. 1*D*). We thus used MD simulations to test whether distinct features of individual isoforms result from static conformational differences, or from a dynamic interface crystallized in different conformations.

We performed 120-ns all-atom MD equilibrium simulations of 4 different EC1-4 homodimers from different Pcdh subfamilies to distinguish the above hypotheses: PcdhγB3, PcdhγB7, Pcdhβ6, and Pcdhα7 (*SI Appendix*, Table S1). Overall, the complexes did not dissociate over the course of simulation (*SI Appendix*, Fig. S1*A* and Fig. 1*B*). While root-mean-squared deviation (RMSD) is not an appropriate metric to determine convergence for the full Pcdh complex due to its large size and the flexibility of the hinge regions between EC repeats (32), overall RMSD remained stable and below ∼8 Å for all structures. Moreover, the RMSD of individual ECs ranged from 2 to 3 Å, showing local equilibrium (*SI Appendix*, Fig. S2). Also, the overall buried surface area (BSA) of each isoform was stable: The average BSA was 4,400 ± 400 Å^2 for PcdhγB3, 4,500 ± 300 Å^2 for PcdhγB7, 4,200 ± 300 Å^2 for Pcdhβ6, and 3,500 ± 200 Å^2 for Pcdhα7.

At an individual EC interaction level, fluctuations in BSA suggest these interactions are weak. Between individual EC domain dimers, which average 700 to 1,300 Å^2 in BSA, fluctuations in BSA can be more than 400 Å^2 over the course of a few nanoseconds (Fig. 1*B* and *SI Appendix*, Table S2). For example, 1 EC1/EC4 dimer of Pcdhα7 fluctuates in BSA from 1,000 Å^2 at the start of the simulation to ∼200 Å^2 at 6 ns then to 1,200 Å^2 around 35 ns (Fig. 1*B*). In another example, our previously published PcdhγB3 structure has a particularly small BSA for the EC2/EC3 interface (15). In our simulations, each EC2/EC3 interface increases in BSA from ∼600 to ∼1,200 Å^2 in the first 6 ns. The presence of HEPES at this interface did not cause the small interface, as structures with less or no HEPES were identical (overall RMSD over 3,210 atoms: 0.890 and 0.785 Å, respectively) (*SI Appendix*, Fig. S3), suggesting this

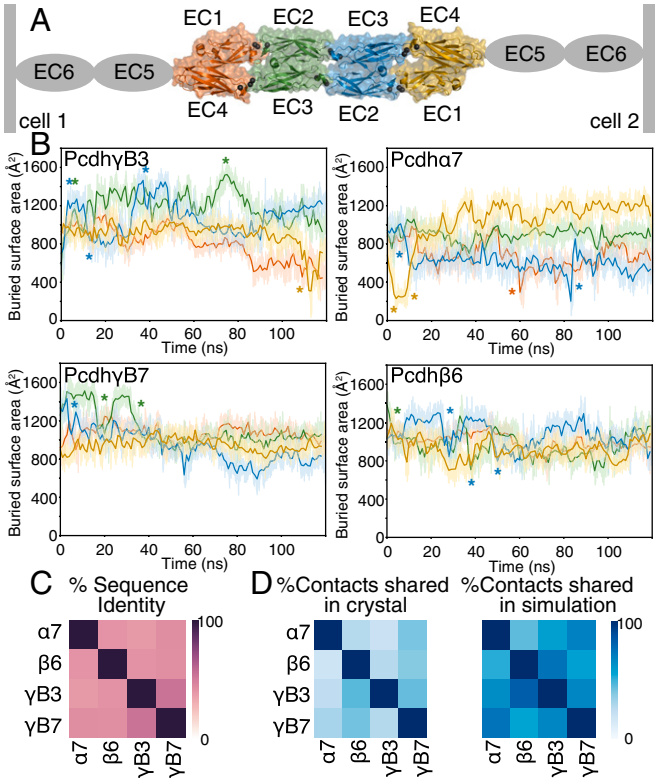

Fig. 1. Clustered Pcdhs have diverse and dynamic interfaces. (*A*) The clustered Pcdh dimer is an antiparallel complex where EC1 of the first protomer interacts with EC4 of the other protomer (orange and yellow) and EC2 interacts with EC3 (blue and green). Each simulation thus provides 2 examples of each type of interaction. (*B*) The BSA of the EC1/EC4 and EC2/EC3 interactions varies throughout the simulations for each simulated isoform (PcdhγB3 EC1-4, PcdhγB7 EC1-4, Pcdhβ6 EC1-4, and Pcdhα7 EC1-5). Instances of rapid interface BSA changes (>400 Å^2 in under 5 ns) are indicated with an asterisk in the corresponding color. (*C*) Percent identity of EC1-4 of the 4 clustered Pcdhs studied here. (*D*) Percentage of interface residue contacts shared between structures (Pcdhα7, 5dzv; Pcdhβ6, 5dzx; PcdhγB3, 5k8r; PcdhγB7, 5szp) and between structures and MD simulations. Each structure shares a mean of 32% of its contacts with other structures, and a mean of 47% of its contacts with other isoforms in simulation.

structure is a conformation sampled by PcdhγB3 in solution. The larger BSA of the equilibrated PcdhγB3 structure is similar to that of other simulated clustered Pcdhs (*SI Appendix*, Fig. S1*B*). In fact, this was generally true of all simulations: the proportion of shared interface sites across the simulations is higher than that across the crystal structures alone (Fig. 1*D*), indicating that simulations can define interface residues representative of a protein family.

Although the MD simulations are not necessarily equilibrated, the fluctuations suggest that if these individual EC interactions were found in isolation they would be low affinity, consistent with the observation that EC1-3 constructs do not dimerize in solution (14, 18). These fluctuations in BSA agree with other simulations that find sharp decreases in BSA as prerequisites to protein complex dissociation (33). Overall, our simulations indicate that each isoform can sample a range of interface conformations, and that the available crystal structures represent only a snapshot of these conformational possibilities. Furthermore, simulations can improve consistency when defining interface residues in a protein family.

Highly Coevolving Residue Pairs Are in Frequent Contact in Simulations. Computational methods based on residue coevolution have been useful in understanding the structure of Pcdhs, predicting the EC1/EC4 interaction and that the *trans* dimer architecture

exists in nonclustered Pcdhs, both findings later confirmed experimentally (15, 16, 18, 34). Coevolving residue pairs identify interprotein contacts (22, 23) but can correspond to positions only in contact in certain conformations (35–37). Given the dynamic nature of the Pcdh interface and the above observation that crystal structures represent only a snapshot of possible conformations, we analyzed how often highly coevolving interface residues in the Pcdhs (15, 18) are in contact across simulations and over time.

In our sequence coevolution analysis, the set of top 200 coevolving pairs includes mostly intramolecular contacts and 15 intermolecular pairs (*SI Appendix*, Fig. S4). We calculated the residue–residue distance of these 15 coevolving intermolecular pairs (Fig. 2*A*) over the course of the MD simulations. Of note, each simulation provides 2 (semi)independent observations for each residue pair due to the 2-fold symmetry of the dimer. In general, pairs with higher coevolution scores are more frequently in contact across more simulations than weaker pairs (Fig. 2*B*), consistent with the observation that lower-scoring pairs are less predictive of physically interacting residues (22, 23).

A closer examination of the top 15 intermolecular pairs reveals diverse trajectories during simulations (Fig. 2*C*). For the most highly coevolving pairs, the residues are in close contact most of the time in most simulations. For example, the 84 to 338 and 123 to 300 pairs (based on PcdhγB3 numbering in *SI Appendix*, Fig. S5) remain at ∼4 Å throughout all but 1 simulation (*SI Appendix*, Fig. S6). For some lower-scoring pairs, the residues are in close proximity in some simulations but further in others, e.g., the 159 to 302 pair residue–residue distances fluctuate between 6 and 18 Å in most simulations but stay consistently close in the Pcdhα7 interface. Other lower scoring pairs fluctuate widely in all simulations and rarely if ever come into contact, such as the 39 to 338 pair.

Model Parameters Reveal Biochemical Interactions Underlying Interface Specificity. Coevolving pairs are calculated using an undirected graphical model, which has parameters for single-site biases and pairwise residue preferences for all sites (24–26). The pairwise residue preferences can capture biochemical relationships between amino acids (38). We asked whether the pairwise residue preferences, termed J_{ij} matrices, can shed light on the biochemical nature of interactions in the interface and inform which residues stay in contact during simulations.

We observe a weak correlation between the pairwise residue preference (J_{ij} value) of a particular residue pair in a particular isoform, and the fraction of the simulation that this pair spends in contact (Pearson's $r = 0.30$, $P < 0.05$, $n = 56$; Fig. 3*B*) The top-scoring pairs, 84–338, 77–371, and 41–342, are a hydrophobic interaction, a charged-aromatic interaction, and a salt bridge, respectively (Fig. 3*A*). These interactions are conserved in their biochemical nature even though the particular residues have undergone substitution. Further down the list of top-scoring intermolecular coevolving pairs, the biochemical nature of the interaction becomes less apparent. For instance, the 124 to 302 pair appears to be an interaction between the hydrophobic position 124 and the charged or polar position 302. These could be due to noise in parameter inference or could indicate positions that are rapidly diversifying and inhabiting repulsive states to achieve specificity. Consistent with this latter hypothesis, these lower-scoring pairs are found predominantly on EC2/EC3, which has been identified as a region of rapid diversifying selection (21). Extreme J_{ij} value residue pairs for EC1/EC4 are concentrated in biochemically similar regions (e.g., 84 to 338 and 77 to 371), while significant residue pair J_{ij} values for EC2/EC3 pairs are broadly distributed in biochemically diverse pairs (*SI Appendix*, Fig. S7).

The dynamics and structural heterogeneity of residue pair interactions in EC2/EC3 may explain why coevolving pairs on this interface have lower scores. Four of the top 15 coevolving pairs cluster in the β4–β5 loop of EC2, which interacts with EC3. During simulations of the PcdhγB3 EC1-4 interface, conformational

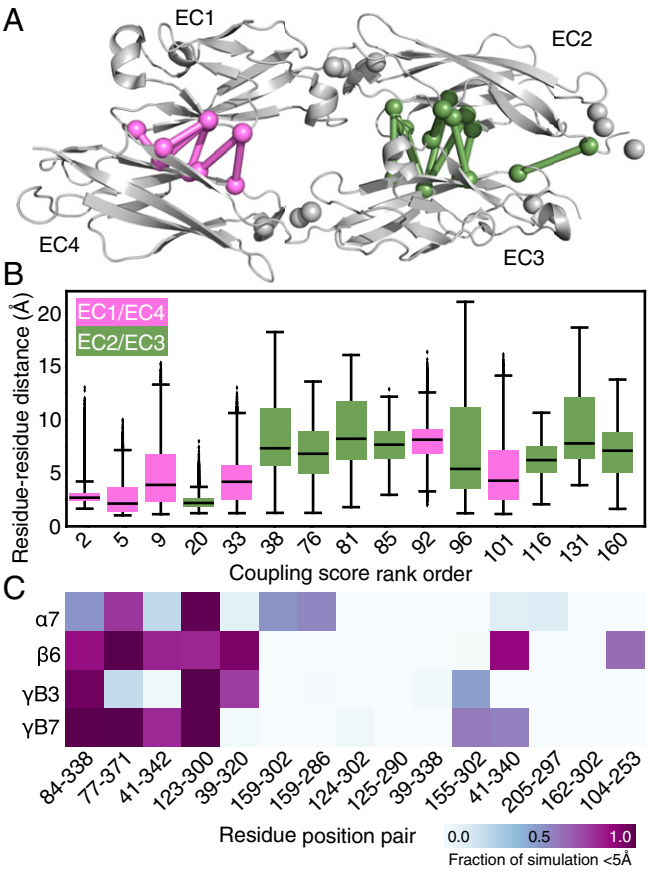

Fig. 2. Evolutionary coupling scores of intermolecular pairs correlate with residue–residue interactions in simulations. (*A*) Clustered Pcdh half-dimer of EC1-2 and EC3-4 showing coevolving residue pairs across the EC1/EC4 (pink) and EC2/EC3 (green) interface. (*B*) Box plots of residue–residue distances for top 15 intermolecular pairs for all 4 simulations. Boxes are colored based on their location on the EC1/EC4 (pink) or EC2/EC3 (green) interface. (*C*) Heatmap of the fraction of time the top 15 intermolecular coevolving pairs are within 5 Å for the 4 simulations.

changes in this loop leads to an increased interface BSA during the first 6 ns of simulation (*SI Appendix*, Fig. S1*B*), including the formation of hydrogen bonds between Y161 and K302, and N155 and T286, and a van der Waals interaction between L156 and M216 (numbering based on *SI Appendix*, Fig. S5 and Fig. 4*A*). The corresponding interresidue distances vary in a coordinated fashion as the loop fluctuates between the disengaged state seen in the structure and an engaged state where the loop interacts directly with EC3 (*SI Appendix*, Fig. S8). The heterogeneity of the EC2 β4–β5 loop is echoed in its diverse conformations in other clustered Pcdh structures (*SI Appendix*, Fig. S9). Pcdhα7 and PcdhγB7 (unlike Pcdhβ6) also use the EC2 β4–β5 loop to interact with EC3 but use a distinct set of amino acid interactions (Fig. 4*B*). These interactions are biochemically diverse, including hydrogen bonds and electrostatic and van der Waals interactions (Fig. 4*C*). For all 3 isoforms, the interactions are dynamic and go through binding and unbinding events (*SI Appendix*, Fig. S8). Isoform differences in how the EC2 β4–β5 loop interacts with EC3 may explain the biochemical diversity of the coevolving residue pairs at this interface and the relative low strength of these couplings.

SEI Describes Pcdh Specificity Distributions. We used evolutionary couplings and simulation data to build a model of clustered Pcdh interaction specificity (*Methods* and *SI Appendix*). We assess the propensity for any 2 Pcdhs to interact by summing the pairwise

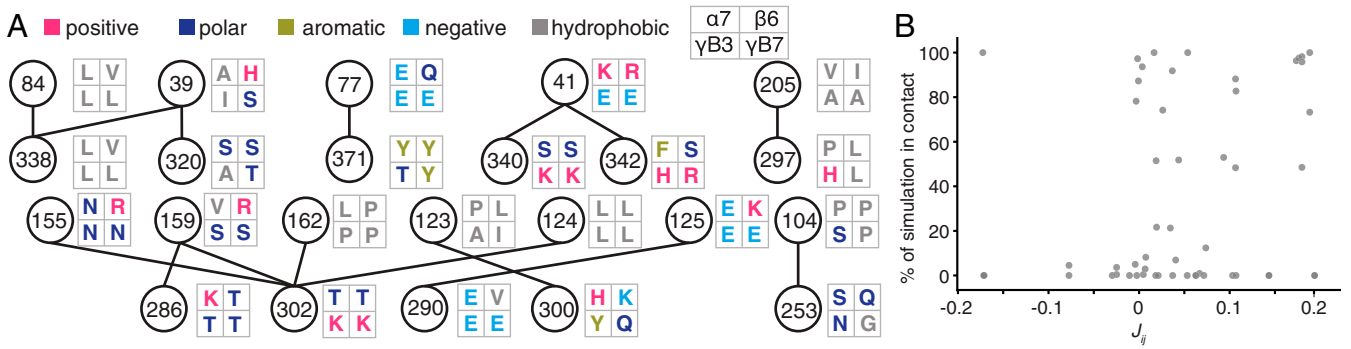

Fig. 3. Coevolving residue pairs reveal biochemical interactions across clustered Pcdh interface. (*A*) The top 15 coevolving pairs and the amino acids found at these positions for the 4 simulations. (*B*) Weak correlation is seen between the pairwise residue preferences (J_{ij} values) and the fraction of the simulation these residues pairs were in contact (within 5 Å).

residue preferences for all interface residue pairs, producing a score that we call the statistical energy of interaction (SEI) (*Methods* and Fig. 5*A*). A higher SEI indicates a higher propensity for interaction.

Our statistical model allows us to compare covariation-based determinants of specificity between all isoforms to discern the relative likelihood of interaction. For the α, β, and γ Pcdh subfamilies in mouse, the SEI of a sequence with itself (a self pairing) is higher than the SEI of a sequence with a different isoform (a nonself pairing; Fig. 5*B*). This is generally consistent with previous cell aggregation experiments in which clustered Pcdhs only form homodimers (12, 13). While these studies observed no nonself interaction in their experiments, our model finds that in some cases the SEI for nonself pairs is as high as for a self pairing, e.g., between β4 and β6. This could be due to particulars of the cell aggregation assay or suggest that some in vivo determinants of specificity are not fully captured by our model.

We compared the contribution of each domain–domain interface to the overall SEI. The nonself Pcdh pairs of the mouse α, γB, and γA subfamilies have a lower SEI in the EC2/EC3 interface than in the EC1/EC4 interface, indicating that the EC2/EC3 interface contributes more to specificity of these isoforms (Fig. 5*C* and *SI Appendix*, Fig. S10). The α subfamily has nearly identical SEI between EC1/EC4 self and nonself interfaces, suggesting this interaction has little discriminatory power in the α subfamily. This finding extends our previous analysis (14) that found that the EC2/EC3 interface tends to contribute more to specificity than the EC1/EC4 interface. The difference in SEI between EC2/EC3 and EC1/EC4 nonself pairs may be due to having mutations between self and nonself EC2/EC3 pairs compared to EC1/EC4 pairs, which negatively correlates with the SEI (*SI Appendix*, Fig. S11).

The SEI model allows us to predict how particular mutations may alter interaction specificity by recalculating SEI using individual coupling terms from the mutant sequence. Previous work has tested chimeric constructs in cell aggregation assays to understand how specificity is encoded in the clustered Pcdh family (14), and our model correctly predicts the phenotype of the majority of these mutants (*SI Appendix*, Fig. S12). On closer inspection of the Pcdhα7/Pcdhα8 pair, the SEI of Pcdhα8 EC1/EC4 nonself pairs are lower than the nonself EC1/EC4 pairs in other α isoforms (*SI Appendix*, Fig. S10). Coincidentally, the Pcdhα7/Pcdhα8 EC2/EC3 SEI is nearly identical to the self SEI scores, suggesting that, within this pair, EC1/EC4 provides more discrimination than EC2/EC3. This agrees with literature results (14) but remains an exception to the trend that EC2/EC3 provide more discriminatory power in the α, γB, and γA subfamilies.

Our model parameters are inferred only from natural sequences, and therefore self pairings, which may bias the model against nonself pairings. To avoid this possible bias, we implemented an iterative pairing algorithm that allows the isoforms to find favorable nonself pairings, if such pairings exist (23, 24) (*SI Appendix*, Fig. S13). The algorithm reproduces self pairings for 74% of all sequences in the alignment, averaged across 5 replicates, after iteration to convergence. This is on par with accuracy of partner detection for other proteins pairs performed by related algorithms (23, 34) and supports the use of our model built from natural sequences. When the EC1/EC4 and EC2/EC3 interactions were paired in isolation, we found that the accuracy of matching

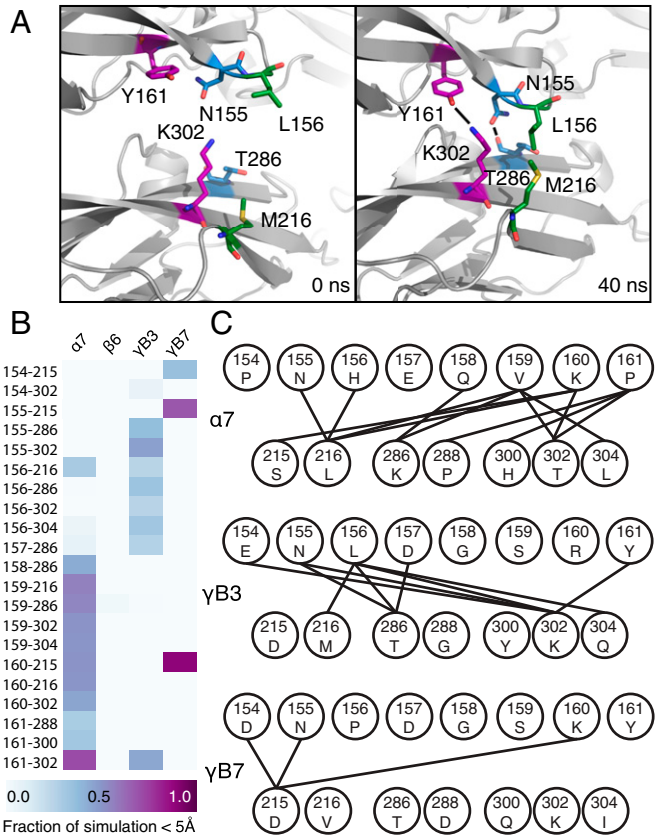

Fig. 4. Structural heterogeneity and sequence diversity of EC2 β4–β5 loop contribute to isoform specificity. (*A*) The EC2 β4–β5 samples conformational space in the PcdhγB3 simulation, forming several interactions with EC3 (numbering based on *SI Appendix*, Fig. S5). (*B*) Residue–residue interactions (shown as a heatmap of the fraction of the simulation pairs within 5 Å) of the EC2 β4–β5 loop (positions 154 to 161) with EC3 (positions 215, 216, 286, 288, 300, 302, and 304) differ between isoforms. (*C*) The differences in interacting residue pairs reflect diverse biochemical interactions of this loop with EC3.

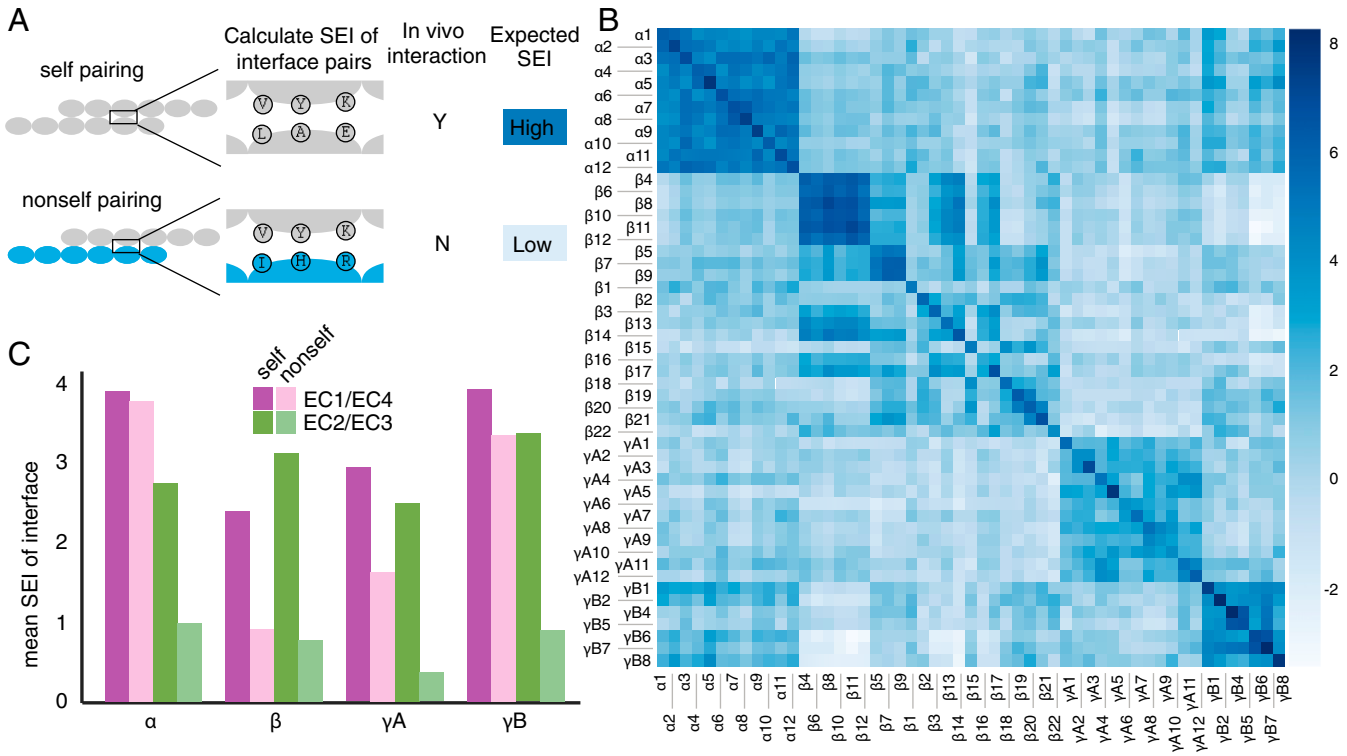

Fig. 5. Statistical energy of interaction supports experimental evidence for self-interaction and divergent roles of EC2/EC3 and EC1/EC4 interfaces. (*A*) SEI for every possible pairing of Pcdhs was calculated by summing the energy contribution of each interacting residue pair at the interface. Based on prior work (27, 28), we expect to observe high statistical energy for self pairings, which interact in vivo (12, 13), and low statistical energy for nonself pairings. (*B*) Statistical energy for all possible combinations of isoforms. (*C*) Mean statistical energy of every self and nonself pairing for the EC2/EC3 and EC1/EC4 interface, measured within all subfamilies of clustered Pcdhs in mouse.

was less than when the whole interface was used, indicating that both interfaces act in combination to achieve full specificity of the interface.

Our model allows us to compute a SEI score for all pairs of clustered Pcdhs that generally agrees with experimental findings about specificity of Pcdh isoforms. Importantly, it also allowed us to dissect contributions of various interface components at an overall and subfamily level. We observed lower mean SEI at EC2/EC3 nonself interfaces than at EC1/EC4 nonself interfaces, indicating that the EC2/EC3 interface tends to be more involved in specificity.

Discussion

The highly specific antiparallel Pcdh interface that forms between neurons is required for many roles in neuronal development. Our results address the determinants of specificity in this interaction by simulating the dynamics of the Pcdh interface and modeling the contributions of each residue pair to interaction specificity.

Our MD simulations of clustered Pcdh dimers reveal that individual EC interactions sample a range of conformations in every isoform. The variations in BSA for individual EC interactions suggest that the individual interactions are weak, and the overall stability of a Pcdh dimer is established by the polyvalent nature of these individual EC interactions. This type of cooperative binding is widespread in biology and plays roles in multisubunit protein machine assembly, signaling at the membrane, and signaling between cells (39–41). The dynamic nature of clustered Pcdh complexes and the cooperativity of individual EC interactions likely play a role in interaction specificity, motivating us to incorporate these simulations in our statistical model of specificity.

The MD simulations allowed us to observe how coevolving residue pairs across a protein interface vary over time. Overall, we found that higher scoring pairs are closer together throughout the simulations and across multiple homologs, demonstrating a relationship between the strength of an interaction and the coupling score of a residue pair. This result extends previous empirical results showing that coevolving pairs are more likely to be close in 3D (22, 23), and that evolutionary couplings can correspond to multiple incompatible conformations (35–37). We also find that residue–residue distances vary between isoforms, suggesting that coevolving pairs may be important for the stability of only a subset of proteins. Thus, our MD results illustrate that coevolving pairs can represent residue interactions that are present in the ensemble of conformations sampled by the collection of homologous proteins that are present in the coevolutionary model. This knowledge could inform further developments to benchmark structure prediction using coevolution data.

By analyzing the correlation between evolutionary couplings and the behavior of residue pairs in simulations, we have shed light on the role of coevolving residue pairs in the specificity of clustered Pcdh interactions. Generally, the most strongly coevolving pairs are found between EC1 and EC4, and these pairs are frequently in contact. The biochemical nature of these interactions is consistent between isoforms, indicating that they serve a conservative role in the Pcdh interactions. Coevolving pairs between EC2/EC3 are less frequently in contact and their biochemical character changes between isoforms, suggesting these interactions are rapidly diversifying and metastable so that they can flexibly occupy new specificity space.

We constructed a model of interaction specificity from sequence data, using residue pairs found to interact in our simulations. We used this model to evaluate pairs of individual EC interfaces, allowing us to determine the relative likelihood of an interaction. We found that the difference in statistical energy of interaction (ΔSEI) supports literature results that the Pcdh interface is specific for self-interaction. Nonself pairings of the EC2/EC3 interface have lower SEI than nonself EC1/EC4 pairings for the α, γA, and γB

subfamilies, indicating that the EC2/EC3 interface has a greater contribution to specificity. There are some differences between subfamilies as noted previously (15, 20), with the β and γA subfamilies having nearly equal contributions to specificity from both interfaces.

The extent to which protein–protein interactions avoid cross talk with paralogs is dependent on the evolutionary consequences of having promiscuous interactions. To avoid spurious signaling, bacterial 2-component systems have strict specificity encoded in a small number of residues (42, 43). Promiscuous intermediates have been derived experimentally for bacterial toxin–antitoxin and PDZ domains (44, 45), although these sequences likely have not been visited evolutionarily. The work presented here suggests a strategy used by clustered Pcdhs to ensure specificity and yet allow new specificities to easily arise through evolution. Small changes in individual EC affinity caused by a small number of mutations can alter the affinity of the whole dimer through the cooperativity of the individual EC interactions. This strategy may explain the pervasiveness of this interface for cell–cell adhesion in nervous system development.

Methods

Detailed procedures for all methods are provided in *SI Appendix*.

Statistical Interaction Energy Model of Clustered Pcdh Specificity. We used evolutionary couplings to build a model of clustered Pcdh interactions. Previous studies used the statistical energy of an evolutionary couplings model to identify interacting histidine kinase-response regulator pairs (27, 28) and to predict the effects of mutations on protein function (31). For our model, only the interface residue pairs determined by our MD approach were used. The interaction energy between 2 sequences (σ^A, σ^B) is the sum of the individual coupling terms (J_{ij}) between the interface residues of the 2 sequences:

$$SEI(\sigma^A, \sigma^B) = \sum_{\substack{\text{interface } (i,j) \\ \text{contacts}}} J_{ij}\left(\sigma_i^A, \sigma_j^B\right).$$

The J_{ij} term is the matrix of pairwise residue preferences for all possible amino acids in positions i and j (31). The change in SEI(σ) is used to predict whether the interaction will become more or less favorable. See *SI Appendix* for further explanation. See Datasets S1–S3 for interface residues, residue pairs, and an alignment of mouse isoforms.

ACKNOWLEDGMENTS. We thank members of the R.G. and D.S.M. laboratories for stimulating discussions on this project, Raul Araya-Secchi and Bennett Vogt for initial work on PcdhγB3, and the Northeastern Collaborative Access Team (NE-CAT) beamline staff and the SBGrid support team for help with crystallographic data collection and analysis. NE-CAT is funded by NIH (P30 GM124165 and S10 RR029205) and the Advanced Photon Source by the US Department of Energy (DE-AC02-06CH11357). Evolutionary couplings analysis was conducted on the Orchestra High Performance Compute Cluster at Harvard Medical School, funded by NIH (National Center for Research Resources 1S10RR028832-01). MD simulations were performed on Texas Advanced Computing Center's Stampede2 supercomputer (Extreme Science and Engineering Discovery Environment MCB140226) and Ohio Supercomputer Center's Owens cluster (PAS1037). J.M.N. was supported by a National Defense Science and Engineering Graduate Fellowship, and A.G.G. by National Science Foundation Graduate Research Fellowship DGE1144152. Support was also provided by NIH (GM106303 [to D.S.M.] and DC015271 [to M.S.]).

1. M. R. Emond, J. D. Jontes, Inhibition of protocadherin-α function results in neuronal death in the developing zebrafish. *Dev. Biol.* **321**, 175–187 (2008).
2. A. M. Garrett, D. Schreiner, M. A. Lobas, J. A. Weiner, γ-Protocadherins control cortical dendrite arborization by regulating the activity of a FAK/PKC/MARCKS signaling pathway. *Neuron* **74**, 269–276 (2012).
3. D. Kostadinov, J. R. Sanes, Protocadherin-dependent dendritic self-avoidance regulates neural connectivity and circuit function. *eLife* **4**, e08964 (2015).
4. J. Ledderose, S. Dieter, M. K. Schwarz, Maturation of postnatally generated olfactory bulb granule cells depends on functional γ-protocadherin expression. *Sci. Rep.* **3**, 1514 (2013).
5. J. L. Lefebvre, D. Kostadinov, W. V. Chen, T. Maniatis, J. R. Sanes, Protocadherins mediate dendritic self-avoidance in the mammalian nervous system. *Nature* **488**, 517–521 (2012).
6. L. Suo, H. Lu, G. Ying, M. R. Capecchi, Q. Wu, Protocadherin clusters and cell adhesion kinase regulate dendrite complexity through Rho GTPase. *J. Mol. Cell Biol.* **4**, 362–376 (2012).
7. X. Wang *et al.*, Gamma protocadherins are required for survival of spinal interneurons. *Neuron* **36**, 843–854 (2002).
8. J. A. Weiner, X. Wang, J. C. Tapia, J. R. Sanes, Gamma protocadherins are required for synaptic development in the spinal cord. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 8–14 (2005).
9. M. J. Molumby, A. B. Keeler, J. A. Weiner, Homophilic protocadherin cell-cell interactions promote dendrite complexity. *Cell Rep.* **15**, 1037–1050 (2016).
10. G. Mountoufaris, D. Canzio, C. L. Nwakeze, W. V. Chen, T. Maniatis, Writing, reading, and translating the clustered protocadherin cell surface recognition code for neural circuit assembly. *Annu. Rev. Cell Dev. Biol.* **34**, 471–493 (2018).
11. J. Brasch *et al.*, Visualization of clustered protocadherin neuronal self-recognition complexes. *Nature* **569**, 280–283 (2019).
12. D. Schreiner, J. A. Weiner, Combinatorial homophilic interaction between γ-protocadherin multimers greatly expands the molecular diversity of cell adhesion. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 14893–14898 (2010).
13. C. A. Thu *et al.*, Single-cell identity generated by combinatorial homophilic interactions between α, β, and γ protocadherins. *Cell* **158**, 1045–1059 (2014).
14. R. Rubinstein *et al.*, Molecular logic of neuronal self-recognition through protocadherin domain interactions. *Cell* **163**, 629–642 (2015).
15. J. M. Nicoludis *et al.*, Antiparallel protocadherin homodimers use distinct affinity- and specificity-mediating regions in cadherin repeats 1-4. *eLife* **5**, e18449 (2016).
16. S. R. Cooper, J. D. Jontes, M. Sotomayor, Structural determinants of adhesion by Protocadherin-19 and implications for its role in epilepsy. *eLife* **5**, 1–22 (2016).
17. X. Peng *et al.*, Affinity capture of polyribosomes followed by RNAseq (ACAPseq), a discovery platform for protein-protein interactions. *eLife* **7**, e40982 (2018).
18. J. M. Nicoludis *et al.*, Structure and sequence analyses of clustered protocadherins reveal antiparallel interactions that mediate homophilic specificity. *Structure* **23**, 2087–2098 (2015).
19. K. M. Goodman *et al.*, Structural basis of diverse homophilic recognition by clustered α- and β-protocadherins. *Neuron* **90**, 709–723 (2016).
20. K. M. Goodman *et al.*, γ-Protocadherin structural diversity and functional implications. *eLife* **5**, e20930 (2016).
21. Q. Wu, Comparative genomics and diversifying selection of the clustered vertebrate protocadherin genes. *Genetics* **169**, 2179–2188 (2005).
22. T. A. Hopf *et al.*, Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3**, e03430 (2014).
23. S. Ovchinnikov, H. Kamisetty, D. Baker, Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030 (2014).
24. D. S. Marks *et al.*, Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766 (2011).
25. F. Morcos *et al.*, Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293–E1301 (2011).
26. J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, J. N. Onuchic, Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 10340–10345 (2012).
27. A.-F. Bitbol, R. S. Dwyer, L. J. Colwell, N. S. Wingreen, Inferring interaction partners from protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12180–12185 (2016).
28. T. Gueudré, C. Baldassi, M. Zamparo, M. Weigt, A. Pagnani, Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12186–12191 (2016).
29. R. R. Cheng, F. Morcos, H. Levine, J. N. Onuchic, Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E563–E571 (2014).
30. R. R. Cheng *et al.*, Connecting the sequence-space of bacterial signaling proteins to phenotypes using coevolutionary landscapes. *Mol. Biol. Evol.* **33**, 3054–3064 (2016).
31. T. A. Hopf *et al.*, Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
32. M. Sotomayor, K. Schulten, The allosteric role of the Ca^{2+} switch in adhesion and elasticity of C-cadherin. *Biophys. J.* **94**, 4621–4633 (2008).
33. L. Zhang, S. Borthakur, M. Buck, Dissociation of a dynamic protein complex studied by all-atom molecular simulations. *Biophys. J.* **110**, 877–886 (2016).
34. J. M. Nicoludis, R. Gaudet, Applications of sequence coevolution in membrane protein biochemistry. *Biochim. Biophys. Acta Biomembr.* **1860**, 895–908 (2018).
35. A. Toth-Petroczy *et al.*, Structured states of disordered proteins from genomic sequences. *Cell* **167**, 158–170.e12 (2016).
36. T. A. Hopf *et al.*, Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607–1621 (2012).
37. F. Morcos, B. Jana, T. Hwa, J. N. Onuchic, Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 20533–20538 (2013).
38. A. Coucke *et al.*, Direct coevolutionary couplings reflect biophysical residue interactions in proteins. *J. Chem. Phys.* **145**, 174102 (2016).
39. M. Mammen, S. K. Choi, G. M. Whitesides, Polyvalent interactions in biological systems: Implications for design and use of multivalent ligands and inhibitors. *Angew. Chem. Int. Ed. Engl.* **37**, 2754–2794 (1998).
40. J. D. Badjić, A. Nelson, S. J. Cantrill, W. B. Turnbull, J. F. Stoddart, Multivalency and cooperativity in supramolecular chemistry. *Acc. Chem. Res.* **38**, 723–732 (2005).
41. A. Whitty, Cooperativity and biological complexity. *Nat. Chem. Biol.* **4**, 435–439 (2008).
42. A. I. Podgornaia, M. T. Laub, Determinants of specificity in two-component signal transduction. *Curr. Opin. Microbiol.* **16**, 156–162 (2013).
43. E. J. Capra, B. S. Perchuk, J. M. Skerker, M. T. Laub, Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. *Cell* **150**, 222–232 (2012).
44. C. D. Aakre *et al.*, Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594–606 (2015).
45. A. S. Raman, K. I. White, R. Ranganathan, Origins of allostery and evolvability in proteins: A case study. *Cell* **166**, 468–480 (2016).